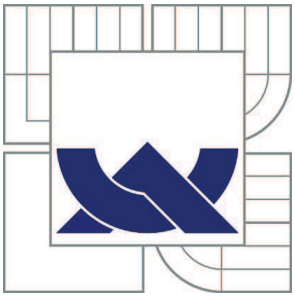
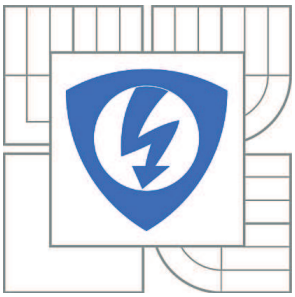


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

ANALÝZA A MODELOVÁNÍ PROVOZU V DATOVÝCH SÍTÍCH

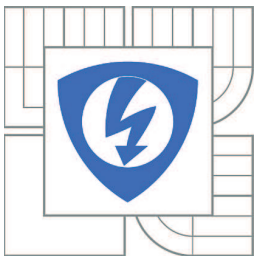
ANALYSIS AND MODELING OF NETWORK DATA TRAFFIC

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. JÁN PAUKEJE
AUTHOR

VEDOUCÍ PRÁCE — Ing. LUKÁŠ RŮČKA
SUPERVISOR

BRNO 2012



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student: Bc. Ján Paukeje ***ID:*** 72989
Ročník: 2 ***Akademický rok:*** 2011/2012

NÁZEV TÉMATU:

Analýza a modelování provozu v datových sítích

POKYNY PRO VYPRACOVÁNÍ:

Prozkoumejte matematické modely využívající zpracování dat ve formě časové řady pro analýzu a modelování síťového provozu. Zaměřte se na popis příchodů jednotlivých datových jednotek, jejich četnost atd. Identifikujte základními postupy využitelné pro analýzu a modelování provozu v datových sítích. Podrobně popište alespoň 2 metody (např. ARIMA, ARCH) pro analýzu a modelování síťového provozu s vysokou variabilitou (provoz protokolu HTTP). Získejte dlouhodobý záznam (alespoň 48 hodin) provozu protokolu HTTP. Navrhněte automatizovatelnou metodu odvození či odhadu parametrů pro matematický popis zachyceného provozu pomocí vybraných metod. Důkladně zdokumentujte použité postupy. Na základě odvozených parametrů a ze znalosti předchozího průběhu provozu namodelujte chování budoucího síťového provozu. Porovnejte výsledné vlastnosti a chování reálného a modelovaného síťového provozu. Dosažené výsledky zdokumentujte.

DOPORUČENÁ LITERATURA:

[1] SAMORODNITSKY Gennady. Long Range Dependence. Foundations and Trends in Stochastic Systems. 2006, č. 3, vyd. 1, s. 163–257.
[2] GERSHENFELD, Neil. The Nature of Mathematical Modeling. 1. vyd. Cambridge: Cambridge University Press, 1999. 356 s. ISBN 978-0521570954.
[3] FAPOJUWO, A., LEE, I. Mathematical Modeling and Characterization of Wireless Network Traffic. Hauppauge: Nova Science Publishers, 2008. 101 s. ISBN 978-1604568691.

Termín zadání: 6.2.2012 ***Termín odevzdání:*** 24.5.2012

Vedoucí práce: Ing. Lukáš Růčka
Konzultanti diplomové práce:

prof. Ing. Kamil Vrba, CSc.
Předseda oborové rady

ANOTÁCIA

Diplomová práca pojednáva o modelovaní sieťovej prevádzky so zameraním na spracovanie pomocou analýzy časových radov. Rozoberá charakteristické vlastnosti sieťovej prevádzky, predovšetkým prevádzky HTTP. Prvé kapitoly sú venované teórii, ktorá popisuje časové rady a ich základné modely, lineárne AR, MA, ARMA, ARIMA a nelineárne ARCH a GARCH. V ďalších kapitolách sú definované pojmy ako sebe-podobnosť a dlhodobá závislosť. Je demonštrované zlyhanie konvenčných modelov pri zachytení týchto špecifických vlastností dátovej prevádzky. Na základe štúdia je v 6. kapitole podrobne popísaný vybraný zmiešaný ARIMA/GARCH model a postup pri stanovení jeho parametrov. V praktickej časti je popísaný postup pri konštrukcii a prispôsobení modelu sieťovej zachytenej prevádzke využívajúc prostredie Matlab. Po stanovení modelu je vykonaná predikcia niekoľkých pozorovaní, ktorá je následne porovnaná so skutočnými hodnotami.

Kľúčové slová: modelovanie sieťovej prevádzky, analýza časových radov, sebe-podobnosť, ARIMA, GARCH, HTTP

ABSTRACT

Theses deals with network traffic modeling focused on elaboration by time series analysis. The nature of network traffic is discussed above all http traffic. First three chapters are theoretical, which describes time series and basic models, linear AR, MA, ARMA, ARIMA and nonlinear ARCH. Other chapters define terms like self-similarity and long range dependence. It is demonstrated a failure of conventional models which cannot capture these specific properties of network data traffic. On the basis of study in chapter 6. is closely described the combined ARIMA/GARCH model and its parameter estimation procedure. Applied part of this theses deals with procedure of estimation and fitting the estimation model to observed network traffic. After an estimation a few future values are predicted on the basis of estimated model. These predicted values are consequently compared with real data.

Keywords: network traffic modeling, time series analysis, self-similarity, ARIMA, GARCH, HTTP

PAUKEJE, J. *Analýza a modelování provozu v datových sítích.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2012. 51 s. Vedoucí diplomové práce Ing. Lukáš Růčka.

Prehlásenie

Prohlašuji, že svou diplomovou práci na téma **Analýza a modelování provozu v datových sítích** jsem vypracoval samostatně pod vedením vedoucího semestrálního projektu a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brne dňa .....

.....
podpis autora

Poďakovanie

Ďakujem vedúcemu práce Ing. Lukášovi Růčkovi za užitočnú metodickú pomoc a cenné rady pri zpracovaní diplomovej práce.

V Brne dňa .....

.....
podpis autora

Obsah

Úvod

S rastúcou integráciou služieb, ktoré poskytujú komunikačné siete, rastú aj nároky. Najjednoduchším spôsobom ako uspokojiť požiadavky je rozširovanie sietí. Tu sa ale dostávame do konfrontácie s rôznymi obmedzeniami, napríklad finančné prostriedky. Ďalšou možnosťou ako uspokojiť nároky a to vo všetkých oblastiach služieb je zvýšenie efektivity. To znamená maximálne využitie dostupných prostriedkov, obzvlášť pre oblasť komunikačných sietí. Zabezpečiť výkon a zaručiť kvalitu v sieťach majú na starosti manažment a optimalizácia.

Najlepším spôsobom ako zefektívniť výkon sietí, je zistiť ako sa správa prevádzka a ako sa prevádzka bude správať v budúcnosti, to znamená , popísať ju a na základe popisu sa ju snažiť predikovať. To znamená, že analýza , modelovanie a predikcia poskytuje informácie, ktoré môžeme využiť pri riadení, optimalizovaní ale aj plánovaní a projektovaní v komunikačných technológiách. Základom pre analýzu a modelovanie sú významné odbory štatistika a matematika, vďaka ktorým je okrem samotnej analýzy možné automatizovať.

Táto práca je zameraná na vybrané metódy analýzy, konkrétne na analýzu a modelovanie časových radov. Tieto metódy sú rozšírené najmä v ekonometrii, svoje uplatnenie si však čoraz viac nachádza aj v ostatných oblastiach. V odvetví komunikácií najmä vďaka spomínanému rapídnemu vývoju, čo spôsobilo zmenu charakteru prevádzky v komunikačných sieťach. Predpokladá sa, že na základe ďalšieho skúmania, sú modely časových radov schopné popísať a predikovať špecifické vlastnosti sieťovej prevádzky lepšie ako konvenčné modely. Nachádzajú tak uplatnenie v oblasti zabezpečenia kvality služieb.

V nasledujúcich kapitolách sa oboznámime so základnými pojmami teórie časových radov, vlastnosťami jednotlivých modelov, postupmi pri analýze a konštrukcii modelov so zameraním na sieťovú prevádzku. Podrobne bude popísaný nami vybraný model zachytenej sieťovej prevádzky a postup pri jeho stanovení.

1 Úvod do analýzy časových radov

1.1 Stochastický proces a časový rad

Stochastický proces (náhodný proces) je množina náhodných veličín $\{X(s, t), s \in S, t \in T\}$, kde S je výberový priestor a T je indexový rad. Pre každé $t \in T$ je $X(., t)$ náhodná veličina definovaná na výberovom priestore S a pre každé $s \in S$ je $X(s, .)$ realizácia stochastického procesu definovaná na indexovom rade T. To znamená, že $X(s, t)$ je náhodná veličina závislá na čase [3, 6].

Stochastické procesy sú zaťažené neurčitosťou, na rozdiel od deterministických procesov, ktoré sa dajú jednoznačne popísať matematickou rovnicou. Stochastický prístup umožňuje popísať výskyt extrémov v časových radoch, čo je veľmi dôležité v praxi takmer vo všetkých odvetviach. V súčasnosti nachádza najčastejšie uplatnenie v ekonomickej oblasti [6, 7] a to vo finančníctve, poisťovníctve.

Časový rad je chronologicky usporiadaná množina dát. Keďže táto práca pojednáva o analýze časových radov so stochastickým prístupom (jednotlivé hodnoty sú generované náhodne), budeme v nasledujúcom texte časový rad $\{y_t, t = 1, 2,\ldots,n\}$ chápať ako konkrétnu realizáciu **stochastického procesu** $\{Y_t\}$. Y_t je náhodná premenná s vlastným rozdelením pravdepodobnosti a každé y_t je jedna hodnota tejto premennej.

Nasledujúca rozprava, ktorá sa zaoberá časovými radami, spolu s kapitolou č. 2, Box-Jenkinsova metodológia, sú výstižne a rozsiahlejšie popísaná v literatúre [6, 7, 9, 10]. Obsahovo je prispôsobená na základe literatúry [12], ktorá v úvode pojednáva o aplikácii konkrétnych modelov v analýze sieťovej prevádzky.

1.2 Základné typy, problémy a charakteristiky časových radov

Časový rad pozostáva z pozorovaní zaznamenaných postupne v čase smerom od minulosti do budúcnosti. Pozorovania sú zaznamenávané v diskrétnych okamihoch avšak intervaly medzi jednotlivými pozorovaniami nemusia byť rovnaké. To prináša problém s voľbou okamihov pozorovania. Najvýhodnejšou možnosťou, s ktorou pracovať, je časový s ekvidistantnými (rovnakými) intervalmi [7] alebo aj **intervalový** časový rad. Je to rad intervalového ukazovateľa, ktorého hodnoty závisia na dĺžke intervalu sledovania. Opakom je **okamihový** časový rad, napríklad príchody správ v sieti, kde sa jednotlivé intervaly medzi príchodmi líšia. Nad okamihovým radom sa pred samotnou analýzou musia vykonať určité operácie: akumulácia hodnôt sledovanej veličiny za dané obdobie alebo spriemerovanie hodnôt v danom časovom intervale. Na základe týchto operácií získame ekvidistantný intervalový časový rad.

Ďalšou vecou, ktorú treba zvážiť pri pozorovaní respektíve pri práci s časovým radom jeho dĺžka, ktorú definuje počet pozorovaní n a nie časové rozpätie medzi prvým a posledným pozorovaním. Problémy s kalendárom (napríklad rôzny počet dní mesiacov).

Máme časový rad $\{y_t, t = 1, 2, \ldots, n\}$, ktorý predstavuje realizáciou stochastického procesu. Každá náhodná veličina sa dá popísať základnými charakteristikami. Skôr ako sa dostaneme k popisu časového radu treba si uvedomiť rozdiel medzi **teoretickými** a **empirickými** charakteristikami [8]. Teoretické charakteristiky sú iba konštanty, ktorých presnú hodnotu nepoznáme, zatiaľ čo empirické charakteristiky sú **odhady** teoretických charakteristík, ktoré sme získali na základe meraní, skúsenosti (grécky empirio).

Základnými **teoretickými charakteristikami** časového radu pre $t = 1, 2, ..., n$ sú:

a) stredná hodnota v čase t $\mu_t = E(y_t)$

(1.1)

b) rozptyl (variancia) v čase t

$$\sigma_t^2 = \text{var}\,(y_t) = \text{D}(y_t) = \text{E}[(y_t - E(y_t))^2] = \text{E}[(y_t - \mu_t)^2] \qquad (1.2)$$

c) autokovariančná funkcia rádu k ($k = 0, \pm 1, \ldots$)

$$\gamma_k(t) = \gamma(t, t-k) = \text{cov}\,(y_t, y_{t-k}) = \text{E}[(y_t - \mu_t)\,(y_{t-k} - \mu_{t-k})], \qquad (1.3)$$

d) autokorelačná funkcia rádu k ($k = 0, \pm 1, \ldots$)

$$\rho_k = \frac{\text{cov}(y_t, y_{t-k})}{\sqrt{\sigma_t^2}\sqrt{\sigma_{t-k}^2}} = \frac{\gamma_k(t)}{\sigma_t \sigma_{t-k}}, \qquad (1.4)$$

Predpona „auto“ špecifikuje, že funkcie popisujú koreláciu, resp. kovarianciu v rámci jedného časového radu. Bez tejto špecifikácie predpony by sa jednalo o závislosť medzi dvoma odlišnými časovými radmi. Autokovariančná a autokorelačná funkcia sú funkcie párne $\gamma_k = \gamma_{-k}$, $\rho_k = \rho_{-k}$, preto sa určujú len pre $k \geq 0$.Tieto funkcie popisujú sériovú závislosť hodnôt časovej rady v dvoch časových okamihoch.

Autokovarianciou sa nazýva stredná hodnota súčinu odpovedajúcich centrovaných veličín, $\gamma_y(t_1, t_2) = E[(y(t_1) - \overline{y}(t_1)).(y(t_2) - \overline{y}(t_2))]$. Ak sú hodnoty v okamihoch t a $t - k$ nezávislé autokovariančná funkcia je rovná 0, môžeme povedať že hodnoty sú nekorelované.

Autokorelačná funkcia ρ_k podáva informáciu o sile lineárnej závislosti medzi veličinami y_t a y_{t-k} . Grafický záznam závislosti ρ_k na k sa nazýva periodogram [6].

Odhady teoretických charakteristík (**empirické charakteristiky**) sú:

a) aritmetický priemer $\overline{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$,

(1.5)

V prípade okamihovej časovej rady pri rôznej vzdialenosti časových intervalov sa používa vážený chronologický priemer :

$$\overline{y} = \frac{\frac{y_1 + \text{y}_2}{2} d_2 + \ldots + \frac{y_{n-1} + \text{y}_\text{n}}{2} d_n}{d_2 + \ldots + d_n},$$

(1.6)

kde d_t, t= 2, …, n, je dĺžka jednotlivých intervalov sledovania [6].

b) empirický rozptyl (variancia) $s_y^2 = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \overline{y}\right)^2 = \frac{1}{n}\sum_{t=1}^{n} y_t^2 - \overline{y}_t^2$,

(1.7)

c) empirická autokovariančná funkcia ($k = 0, 1, \ldots$)

$$c_k = \frac{1}{n-k}\sum_{t=1}^{n-k}\left(y_t - \overline{y}\right)\left(y_{t+k} - \overline{y}\right) = \frac{1}{n-k}\sum_{t=1}^{n-k} y_t y_{t+k} - \overline{y}^2,$$

(1.8)

d) empirická autokorelačná funkcia (k =0, 1,…)

$$r_k = \frac{c_k}{s_y^2} = \frac{c_k}{c_0}$$

(1.9)

Keďže sa jedná o odhady (empirické charakteristiky), ich výpočet má zmysel pre časové rady s dostatočnou dĺžkou n ($n > 50$, $k \leq n/4$) [7].

1.3 Základné prístupy k analýze a modelovaniu časových radov

Analýza a zostavenie samotného modelu je zložitý proces, ktorý závisí na množstve faktorov: type dát, dĺžke časového radu, účelu analýzy, výpočtovej technike, software a skúsenostiach analytika. Táto časť práce v stručnosti popisuje základné prístupy k analýze časových radov. Ich znalosť je dôležitá pri selektovaní najvhodnejšej metódy, čo samozrejme šetrí čas a prináša najlepšie výsledky.

Dekompozícia časového radu

Princíp spočíva v rozložení časovej rady na štyri základné zložky:

1) Trendová zložka (T_t) vyjadruje dlhodobú tendenciu vývoja skúmaného javu.
2) Cyklická zložka (C_t) vyjadruje kolísanie okolo trendu v ktorom sa striedajú fázy rastu a poklesu .
3) Sezónna zložka (S_t), tvorí ju pravidelné kolísanie okolo trendu v dôsledku známych udalostí.
4) **Nesystematická zložka** (a_t) označovaná aj ako reziduálna zložka. Túto zložku tvoria náhodné a nesystematické výkyvy ale aj chyby merania. V dekompozícií sa používa označenie I_t [6], v regresných modeloch ,ARIMA modeloch a v tomto texte pre zjednodušenie sa označuje ako a_t. S touto zložkou dekompozičné metódy nepracujú. Táto zložka bude podrobnejšie popísaná v ďalších kapitolách.

Samotný rozklad sa robí pomocou dvoch spôsobov:

a) Aditívny $y_t = T_t + S_t + C_t + a_t$, $t = 1, 2, ..., n$
b) Multiplikatívny $y_t = T_t \times S_t \times C \times a_t$, $t = 1, 2, ..., n$, pri tejto dekompozícií je trendová

zložka meraná v rovnakých jednotkách ako y_t a ostatné zložky sú bezrozmerné veličiny.

Medzi ďalšie prístupy k analýze patria **lineárne kauzálne (faktorové) modely** často používané v ekonometrii a **spektrálna analýza časových radov**, ktorá využíva Fourierovu analýzu. S týmito modelmi spolu s **dekompozíciou časového radu** sa práca nebude ďalej zaoberať.

Box-Jenkinsova metodológia

V rámci tohto prístupu sa predpokladá stacionárny časový rad. Základom pri analýze a vytvorení modelu je **reziduálna zložka** a_t. Práca s touto zložkou umožňuje vytvárať flexibilné modely časových radov, ktoré sú schopné adaptovať sa vysokú variabilitu v ich priebehu. Využívajú sa predovšetkým metódy korelačnej analýzy, ktorá popisuje závislosť medzi jednotlivými pozorovaniami časového radu. Stacionarita a Box-Jenkinsova metodológia budú popísané podrobnejšie v nasledujúcej kapitole. Podrobnejšie z dôvodu využitia metodológie a jej modelov pri analýze sieťovej prevádzky. Skutočnosť, že tieto modely nachádzajú uplatnenie v oblasti analýzy sieťovej prevádzky je preberaná v [12].

2 Box-Jenkinsova metodológia

2.1 Stacionarita

Aby bolo možné v praxi stochastický proces analyzovať musí spĺňať podmienku stacionarity. Analýza nestacionárneho procesu je aj napriek tomu možná. A to v prípade ak by sme nestacionárny proces diferencovali na čiastkové procesy, ktoré už sú stacionárne alebo kvazistacionárne [3]. Túto možnosť zatiaľ nebudeme brať do úvahy a budeme uvažovať iba stacionárny stochastický proces.

Striktná stacionarita (alebo stacionarita v užšom zmysle), predpokladá, že rozdelenie náhodného procesu je invariantné voči posunom v čase, t.j. jeho úplný štatistický popis sa nemení. Pretože túto podmienku nie je jednoduché v praxi dodržať bol zavedený pojem slabá (kovariančná) stacionarita.

Slabá stacionarita (tiež stacionarita v širšom zmysle) je naopak menej obmedzujúca a požaduje aby mal príslušný stochastický proces konštantné momenty druhého rádu, t.j. strednú hodnotu, rozptyl. Kovariančná a korelačná funkcia slabo stacionárneho procesu závisí iba na časovej vzdialenosti (posunutí) náhodných veličín a nie na umiestnení v časovej rade [6, 7]. Časovú **vzdialenosť *k*** v spojení s autokorelačnou funkciu nazývame rádom a má významnú úlohu pri určení správneho modelu. Okrem iného, je obmedzenie sa na kovariančne stacionárny (časovo invariantný) stochastický proces výhodné aj z dôvodu zjednodušenia vzťahov teoretických charakteristík. To znamená, že pre $t = 0,1,...n$ platí:

a) stredná hodnota $$\mu_t = \mathrm{E}(y_t) = \mu \tag{2.1}$$

b) rozptyl $$\sigma_t^2 = \mathrm{var}(y_t) = \mathrm{D}(y_t) = \mathrm{E}[(y_t - \mu_t)^2] = \sigma^2 \tag{2.2}$$

c) autokovariančná funkcia rádu k ($k = 0, \pm 1, \ldots$)

$$\gamma_k = \gamma(t, t-k) = \mathrm{cov}\,(y_t, y_{t-k}) = \mathrm{E}[(y_t - \mu)\,(y_{t-k} - \mu)], \tag{2.3}$$

d) autokorelačná funkcia rádu k ($k = 0, \pm 1, \ldots$)

$$\rho_k = \frac{\mathrm{cov}(y_t, y_{t-k})}{\sigma^2} = \frac{\gamma_k}{\gamma_0}, \tag{2.4}$$

Ak by sme sa v praxi rozhodli vizuálne posúdiť, či je rad stacionárny. Platí, že grafické zobrazenie stacionárneho procesu (v našom prípade časového radu) sa zdá byť ploché, t.j. bez klesajúceho, či stúpajúceho trendu. A bez periodických výkyvov alebo inak sezónnosti

2.2 Autokorelačná funkcia (ACF), parciálna autokorelačná funkcia(PACF)

Vzťahy pre výpočet teoretickej autokorelačnej funkcie rádu k (1.4) a autokorelačnej funkcie rádu k stacionárneho časového radu (2.4) ako aj jej odhad (1.9) boli uvedené v predchádzajúcej kapitole. Pre úplnosť, hodnoty tejto funkcie sa nazývajú ako autokorelácia rádu k . Pre obor hodnôt autokorelačnej funkcie rádu k funkcie vždy platí

$$|\rho_k| \leq 1 \text{ pre všetky } k > 0 \tag{2.5}$$

pre $k = 0$ je autokorelačná funkcia rovná 1, $\rho_0 = 1$.

Pozn.: *V nasledujúcom texte budeme pre autokorelačnú funkciu používať skratku ACF z anglického názvu Autocorrelation Function a obdobne skratku PACF pre parciálnu autokorelačnú funkciu.*

Priebeh ACF je významným faktorom pri výbere modelu pre danú časovú radu. V podstate sa snažíme určiť takú hodnotu vzdialenosti resp. rádu *k,* pri ktorej začína byť teoretická ACF nulová, $k = k_0$. Prípadne môžeme zistiť, že takáto hodnota vôbec neexistuje. Z poznatkov o ACF môžeme usúdiť, že tento bod predstavuje akúsi hraničnú vzdialenosť, pri ktorej hodnoty časovej rady prestávajú byť na sebe závislé. Nakoľko teoretické hodnoty ρ_k nepoznáme, jedinou možnosťou ako overiť túto hypotézu je výpočet jej odhadu (1.9)z pozorovaní. Funkcia r_k je **odhadom** funkcie ρ_k, nazývame ju aj výberová autokorelácia s oneskorením *k.* Pre úplnosť je postup výpočtu r_k rozpísaný vo vzťahu (2.6).

$$r_k = \hat{\rho}_k = \frac{\sum_{t=k+1}^{n} \left(y_t - \bar{y}\right)\left(y_{t-k} - \bar{y}\right)}{\sum_{t=1}^{n} \left(y_t - \bar{y}\right)^2}, \quad k = 1, 2, ..., n-1. \tag{2.6}$$

Testuje sa teda nulová hypotéza H_0: $\rho_1 = \rho_2 = ... = \rho_k = 0$ proti hypotéze H_1: non H_0. Pri rozhodovaní o pravdivosti jednej alebo druhej hypotézy porovnávame vypočítanú hodnotu $|r_k|$ s dvojnásobkom smerodajnej odchýlky funkcie r_k , teda s hodnotou $2\sigma(r_k)$. Voľba hodnoty dvojnásobku $\sigma(r_k)$ približne zodpovedá 5% hladine významnosti testu. Pre stanovenie smerodajnej odchýlky $\sigma(r_k)$ sa používa Bartlettova aproximácia (2.7) [7]. Ak je $\rho_k = 0$ pre $k>k_0$, potom platí

$$\sigma\left(r_k\right) \approx \sqrt{\frac{1}{n}\left(1 + 2\sum_{j=1}^{k_0} r_j^2\right)}, \tag{2.7}$$

Nesmieme zabúdať na fakt, že bod *k* spĺňajúci hypotézu nemusí vôbec existovať.

Korelácia medzi náhodnými veličinami y_t a y_{t-k} môže byť spôsobená aj ich koreláciou s veličinami y_{t-1}, y_{t-2} ,... y_{t-k+1} t.j. veličinami ležiacich medzi nimi [6]. Parciálna autokorelačná funkcia (PACF) tento fakt zohľadňuje a vyjadruje závislosť medzi y_t a y_{t-k} "očistenú" o vplyv veličín y_{t-1}, y_{t-2} ,... y_{t-k+1} . PACF s oneskorením *k* popisuje parciálny regresný koeficient ϕ_{kk} v autoregresii *k*-teho rádu

$$y_t = \phi_{k1} y_{t-1} + \phi_{k2} y_{t-2} + ... + \phi_{kk} y_{t-k} + e_t \tag{2.8}$$

kde veličina e_t nekoreluje s veličinami y_{t-j}, j = 1, 2, ...,*k*. Zo vzťahu autoregresie (2.8)[6] je zjavné, že regresné koeficient ϕ_{k1}, ϕ_{k2}, ... ϕ_{kk} môžeme chápať ako váhy, ktorými sú váhované hodnoty časového radu. Odhady f_{kk} PACF ρ_{kk} sa vypočítajú podľa Durbinovho rekurzívneho vzťahu (2.9)(2.10) [6].

$$f_{kk} = \hat{\rho}_{kk} = \frac{r_k - \sum_{j=1}^{k-1} f_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} f_{k-1,j} r_j}, \text{ pre } \forall\ k > 1, \tag{2.9}$$

$$\text{kde} \quad f_{kj} = f_{k-1,j} - f_{kk} f_{k-1,k-j}, \text{ pre } j = 1, 2, \ldots, k-1. \tag{2.10}$$

Podobne ako u ACF aj u PACF je pre výber vhodného modelu dôležité overiť hypotézu o existencii identifikačného bodu k_0 a pokiaľ existuje zistiť jeho hodnotu.

Postupuje sa analogicky ako v prípade ACF. Smerodajná odchýlka f_{kk} , sa počíta pomocou Quenouilleovej aproximácie [7]. Ak je $\rho_{kk} = 0$, potom platí

$$\sigma(f_{kk}) \approx \sqrt{\frac{1}{n}}, k > k_0 \tag{2.11}$$

2.2.1 Autokorelácia reziduí

Reziduálnu zložku s označením a_t, $t = 1, 2, ..., n$ predstavuje **biely šum** s normálnym rozdelením N(0, σ_a^2). To znamená, že ide o stacionárny stochastický proces tzv. IID proces (proces s nezávislým rozdelením pravdepodobnosti), v ktorom a_t, $t = 1, 2, ..., n$ sú nekorelované premenné s obvykle nulovou strednou hodnotu a s konštantným rozptylom σ_a^2. Autokovariančná funkcia takéhoto procesu je taktiež rovná nule. Zmienené vlastnosti sú matematicky vyjadrené podmienkami: (2.21), (2.22), (2.23). Špeciálnym prípadom je gaussovský šum, proces s normálnym rozdelením N(0, $\sigma_a^2 = 1$). V modeloch reziduálna zložka reprezentuje napríklad chyby v meraní, zlú voľbu regresného modelu a rôzne iné vplyvy. V odbornej literatúre sa pre túto zložku používa rôzne označenie , najčastejšie však epsilon, v tejto práci budeme z dôvodu prehľadnosti využívať označenie a_t.

Reziduálna autokorelačná funkcia

Obdobne, ako autokorelačná funkcia, aj reziduálna autokorelačná funkcia vyjadruje mieru lineárnej závislosti časovo oneskorených veličín a_t a a_{t-k}. Autokorelácia reziduí je definovaná vzťahom

$$r_k = \hat{\rho}_k = \frac{\sum_{t=k+1}^{T} \hat{a}_t \hat{a}_{t-k}}{\sum_{t=1}^{T} \hat{a}_t^2} \tag{2.12}$$

Pomocou reziduálnych (nesystematických) zložiek a_t môžeme modelovať jednoduchý typ autokorelácie

$$y_t = a_t + \rho_1 a_{t-1}. \tag{2.13}$$

Parameter ρ_1 nadobúda hodnoty z intervalu ⟨-1, 1⟩, viď. popis oboru autokorelačnej funkcie (2.5). Hodnoty blízke 1 napovedajú, že ide o pozitívnu koreláciu. Naopak hodnoty blízke -1 signifikujú zápornú koreláciu. V prípade nekorelovanosti nadobúda reziduálna autokorelačná funkcia, respektíve parameter ρ_1, hodnotu rovnú 0. Vzťah (2.13) popisuje proces kĺzavých súčtov 1. rádu (MA(1), preto ρ_1).

Durbin–Watsonov test

Pre overenie nekorelovanosti nesystematickej zložky sa používa Durbin-Watsonov test [6]. Testujeme pomocou prvého koeficientu autokorelácie ρ_1. Ako nulovú hypotézu považujeme rovnosť

H_0: $\rho_1 = 0$ autokorelácia neexistuje, t.j. cov(a_t, a_{t-1}) = 0.

H_1: $\rho_1 \neq 0$ zamietnutie nulovej hypotézy, autokorelácia existuje.

Testovacou štatistikou je Durbin–Watsonovo kritérium, ktoré má tvar

$$DW = \frac{\sum_{t=2}^{T} (\hat{a}_t - \hat{a}_{t-k})^2}{\sum_{t=1}^{T} \hat{a}_t^2} \tag{2.14}$$

a nadobúda hodnoty z intervalu $\langle 0,\ 4\rangle$. Po aproximácií $DW = 2(1-\hat{\rho})$ vzťahu (2.14) dostaneme nám známy vzťah $\hat{\rho}_1 = \frac{\sum_{t=2}^{T} \hat{a}_t \hat{a}_{t-1}}{\sum_{t=1}^{T} \hat{a}_t^{\,2}}$ (viď. autokorelačnú funkciu reziduí (2.12)).

Po dosadení a na základe porovnania DW a $\hat{\rho}_1$ môžeme usúdiť, že pri malých hodnotách DW vykazuje rad pozitívnu autokoreláciu (DW=0, $\hat{\rho}_1 = 1$). Veľké hodnoty DW identifikujú zápornú autokoreláciu (DW=4, $\hat{\rho}_1 = -1$). V prípade DW=2 je $\hat{\rho}_1 = 0$, ide o jasnú nekorelovanosť. Správne rozhodnutie o zamietnutí alebo nezamietnutí testovanej hypotézy na 5% hladine významnosti, prípadne nutnosť zvýšenia T a opakovanie testu (2.14), vyžaduje určenie kritických hodnôt, ktoré sú uvedené v prílohe literatúry [6].

2.3 Lineárny proces a lineárne modely

Lineárny proces môžeme definovať ako nekonečný rad:

$$y_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots, \tag{2.15}$$

kde ψ_i sú vzájomne **nekorelované parametre** a zložky a_t reprezentujú biely šum s nulovou strednou hodnotou a rozptylom σ_a^2. Pri zápise lineárneho procesu sa často používa **operátor spätného posunutia** B, pre ktorý platí $\mathrm{B}^j y_t = y_{t-j}$ to isté platí pre zložku a_t, ktorá bola spomenutá pri pojednávaní o dekompozičných metódach, $\mathrm{B}^j a_t = a_{t-j}$. Pomocou operátora B sa vzťah (2.15) dá zapísať v tvare $y_t = \psi(\mathrm{B})\, a_t$, kde

$$\psi(\mathrm{B}) = 1 + \psi_1 \mathrm{B} + \psi_2 \mathrm{B}^2 + \ldots = 1 + \sum_{j=1}^{\infty} \psi_j \mathrm{B}^j . \tag{2.16}$$

Podmienka stacionarity: Lineárny proces existuje, ak platí, že ψ(B) konverguje pre |B|≤1, B je obyčajná číselná premenná. Táto podmienka zaručuje stacionaritu lineárneho a nulovosť jeho strednej hodnoty E(y_t) = 0.

Prax vyžaduje, aby sa súčasná hodnota y_t lineárneho procesu dala vyjadriť pomocou hodnôt predchádzajúcich y_{t-1}, y_{t-2} ,... a súčasnej hodnoty bieleho šumu a_t v tvare

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + a_t \tag{2.17}$$

a pomocou operátoru posunutia B môžeme (2.17) zapísať $\phi(\mathrm{B})\, y_t = a_t$ kde

$$\phi(\mathrm{B}) = 1 - \phi_1 \mathrm{B} - \phi_2 \mathrm{B} - \ldots = 1 - \sum_{j=1}^{\infty} \phi_j \mathrm{B}^j . \tag{2.18}$$

Lineárne procesy, ktoré sa dajú takto vyjadriť sa nazývajú **invertibilné**. **Podmienka invertibility:** Lineárny proces je invertibilný, ak ϕ (B) konverguje pre |B|≤1. Vzťah medzi parametrami ψ_j a ϕ_j získame z podmienky ψ(B) ϕ(B)=1. Pri výpočte pracujeme s ψ(B) a ϕ(B) ako s mocninovými radami. Po dosadení(2.17) a (2.18) dostaneme

$$1 + (\psi_1 - \phi_1)\mathrm{B} + (\psi_2 - \phi_1\psi_1 - \phi_2)\mathrm{B} + \ldots = 1, \tag{2.19}$$

po úprave (2.19) získame $\psi_1 = \phi_1$, $\psi_2 - \phi_1\psi_1 - \phi_2 = 0$, …analogicky.

V Box-Jenkinsovej metodológií majú praktický význam procesy s konečný, čo najmenším počtom nenulových parametrov. Parametre sa volia tak aby vyhovovali podmienkam pre stacionaritu a invertibilitu.

2.3.1 Proces kĺzavých súčtov MA(q)

Proces kĺzavých súčtov je často označovaný ako proces kĺzavých priemerov (Moving Average), v odbornej terminológií sa používa skratka MA(q).

$$y_t = a_t + \theta_1 a_{t-1} + ... + \theta_q a_{t-q} = \theta_q(\mathrm{B})\, a_t \tag{2.20}$$

Vzťah (2.20) popisuje proces kĺzavých súčtov rádu q. Mocninová rada $\theta_q(\mathrm{B}) = 1 + \sum_{j=1}^{q} \theta_j \mathrm{B}^j$ predstavuje operátor kĺzavých súčtov, pri analýze odhadujeme parametre procesu θ_1, θ_2,…, θ_q , ktoré sú reálne čísla.

Pre zložky bieleho šumu a_t platí:

$$\mathrm{E}(a_t) = 0 \text{, pre } \forall\, t, \tag{2.21}$$

$$\mathrm{D}(a_t) = \mathrm{E}(a_t^2) = \sigma_a^2 > 0, \tag{2.22}$$

$$\mathrm{cov}(a_t, a_s) = \mathrm{E}(a_t a_s) = 0 \text{, pre } t \neq s \tag{2.23}$$

Samotný proces MA(q) má tieto vlastnosti [6, 7]:

1) Je vždy stacionárny, θ_q(B) konverguje pre ľubovoľné hodnoty parametrov.
2) Stredná hodnota je nulová E(y_t) = 0
3) Pre varianciu (rozptyl) platí: $\mathrm{var}(y_t) = \sigma_a^2 \left(1 + \sum_{j=1}^{q} \theta_j^2\right)$
4) ACF má tvar $\rho_k = \dfrac{\theta_k + \theta_1\theta_{k+1} + ... + \theta_{q-k}\theta_q}{\left(1 + \sum_{j=1}^{q} \theta_j^2\right)}$. Identifikačný bod $k_0 = q$.
5) Parciálna autokorelačná funkcia f_{kk} procesu MA(q) nemá identifikačný bod.
6) Proces MA(q) je invertibilný, ak všetky jeho nulové body, korene polynómu θ_q(B) = 0, ležia vo vnútri jednotkovej kružnice [6]. To znamená, že pre korene tohto polynómu musí platiť $|\theta_q| < 1$.

Z uvedených vlastností je dôležité zapamätať si 1) a 6). Z týchto vlastností vyplýva, že stacionarita MA(q) je zaručená. Naopak invertibilitu musíme pri stanovení modelu otestovať.

2.3.2 Autoregresný proces AR(p)

Autoregresný proces rádu p označovaný ako AR(p) je definovaný tvarom

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + a_t . \tag{2.24}$$

Proces AR(1) má len prvý parameter ϕ_1. Ten môžeme chápať ako indikátor pamäti procesu, čím je jeho absolútna hodnota bližšia hodnote 1, tým bude pamäť dlhšia, a naopak, čím bude ϕ_1 bližší nule, tým bude pamäť kratšia. Ak je rovný nule, ACF je nulová , proces nemá pamäť a tým pádom ide iba o samotný proces šumu a_t.

S použitím operátora spätného posunutia môžeme napísať $\phi(\mathrm{B})\, y_t = a_t$, kde

$$\phi(\mathrm{B}) = 1 - \phi_1 \mathrm{B} - \phi_2 \mathrm{B}^2 - ... - \phi_p \mathrm{B}^p$$

je autoregresný operátor.

Základné vlastnosti procesu AR(p)[6, 7]:

1) AR(p) je stacionárny, ak všetky korene polynómu ϕ_p(B) ležia vo vnútri jednotkovej kružnice.
2) Pre strednú hodnotu platí E(y_t) = 0

3) Pre rozptyl platí

$$\mathrm{var}\left(y_t\right) = \frac{\sigma_a^2}{1-\phi_1\rho_1-\phi_2\rho_2-\ldots-\phi_p\rho_p}. \tag{2.25}$$

4) ACF procesu AR(p) spĺňa sústavu diferenčných rovníc

$$\rho_k = \phi_1\rho_{k-1}+\phi_1\rho_{k-2}+\ldots+\phi_p\rho_{k-p}, \text{ pre } k>0. \tag{2.26}$$

ACF je lineárna kombinácia klesajúcich geometrických postupností sínusoíd rôznych frekvencií s geometricky klesajúcimi amplitúdami [7].

5) PACF f_{kk} má identifikačný bod $k_0 = p$, z čoho vyplýva, že $f_{kk} = 0$ pre $k > p$.
6) Proces AR(p) je invertibilný pre ľubovoľné hodnoty parametrov.

2.3.3 Zmiešaný proces ARMA(p, q)

Kvôli rozličným vlastnostiam jednotlivých procesov/modelov sa v praxi používajú zmiešané modely. Zmiešaný model procesov AR(p) a MA(q) sa označuje ARMA (p, q).

Proces rádu p a q je definovaný vzťahom

$$y_t = \phi_1 y_{t-1}+\phi_2 y_{t-2}+\ldots+\phi_p y_{t-p}+a_t+\theta_1 a_{t-1}+\theta_2 a_{t-2}+\ldots+\theta_q a_{t-q} \tag{2.27}$$

alebo pomocou operátora spätného posunutia

$$\phi_p\left(\mathrm{B}\right) y_t = \theta_q\left(\mathrm{B}\right) a_t$$

kde ϕ (B) je autoregresný operátor a θ (B) operátor kĺzavých súčtov. Vlastnosti zmiešaného proces sa dajú odvodiť z vlastností čiastkových procesov AR(p) a MA(q).

1) ARMA (p, q) je stacionárny, ak je stacionárny proces AR(p). Teda platí $|\varphi 1| < 1$
2) Stredná hodnota stacionárneho ARMA (p, q) je nulová.
3) ACF vyhovuje sústave diferenčných rovníc (2.26) pre $k > q$. Nemá identifikačný bod a po prvých $q - p$ hodnotách predstavuje lineárnu kombináciu klesajúcich geometrických postupností a sínusoíd rôznych frekvencií s klesajúcimi amplitúdami.
4) To isté ako v bode 3, platí aj pre PACF tohto procesu.
5) ARMA proces je invertibilný ak je invertibilný proces MA(q) $|\phi_1| < 1$.

V praxi sa najčastejšie stretneme s modelom ARMA(1,1). Z definičného vzťahu (2.27) je zrejmé, že ak pri procese ARMA(1,1) $\phi_1 = 0$, proces sa redukuje na proces MA(1) a naopak, ak $\theta_1 = 0$, ARMA(1,1) sa redukuje na proces AR(1). Proces ARMA(1,1) je invertibilný ak $|\theta_1| < 1$ a stacionárny ak $|\phi_1| < 1$

2.3.4 Integrované procesy ARIMA(p, d, q)

Tento model je určený pre popis časových radov, ktoré nemusia byť stacionárne. Je však nutné aby sa dal previesť na stacionárny. To sa realizuje diferencovaním na čiastkové rady, ktoré už sú stacionárne. Zmiešaný integrovaný model ARIMA (p, d, q) sa popisuje vzťahom

$$\phi_p\left(\mathrm{B}\right)\Delta^d y_t = \theta_q\left(\mathrm{B}\right) a_t \tag{2.28}$$

$\Delta^d y_t$ predstavuje časový rad skonštruovaný diferenciáciou pôvodného časového radu a d vyjadruje rád tejto diferenciácie.Δ =(1–B) je diferenčný operátor. Môžeme napísať:

$$\Delta^d y_t = y_t\left(1-\mathrm{B}\right)^d \tag{2.29}$$

Ak by sme použili diferencovanie napríklad rádu d = 1, 2,… dostali by sme pre d = 1 $\Delta y_t = y_t - y_{t-1}$, a pre d = 2, $\Delta^2 y_t = y_t - 2\,y_{t-1} + y_{t-2}$ a analogicky.

Zostavenie takéhoto modelu sa realizuje v dvoch krokoch:

1) Najskôr sa pôvodný časový rad y_t prevedie diferencovaním, respektíve vhodnou transformáciou pred diferencovaním, na stacionárny rad $\Delta^d y_t$. Diferencovaním sa pôvodný časový rad skráti.

2) Ako ďalší krok sa aplikuje zmiešaný model ARMA (p, q)

Otázkou zostáva výber vhodného rádu diferenciácie t.j. parametru d. Existuje niekoľko spôsobov:

a) Vizuálne posúdenie stacionarity časového radu y_t a diferencovaných radov $\Delta^d y_t$, $d = 1, 2, \ldots$.
b) Skúmanie odhadov ACF r_k pre rady y_t a $\Delta^d y_t$. Ak hodnoty r_k klesajú pomaly, vyskytuje sa dlhodobá závislosť, je nutné diferencovať znovu.
c) Ďalšou možnosťou je skúmanie odhadov rozptylu pre pôvodnú časový rad a diferencované časové rady. Výsledkom je voľba hodnoty parametra d, ktorá so sebou prináša najmenšiu hodnotu odhadu rozptylu.

Čo sa týka transformácií pre diferencovanie časového radu, najčastejšie sa používa Box-Jenkinsova transformácia v tvare

$$y_t^{(\lambda)} = \begin{cases} y_t^{\lambda} & \text{pre } \lambda \neq 0 \\ \log y_t & \text{pre } \lambda = 0 \end{cases}. \tag{2.30}$$

Pričom hodnota parametra λ sa určuje graficky [10]. Časový rad sa rozdelí na kratšie úseky v každom úseku sa určí aritmetický priemer hodnôt m a rozdiel medzi maximálnou a minimálnou hodnotou r. Zostrojí sa graf závislosti r na m. Pre výber optimálneho λ potom platí:

1) Ak je hodnota r konštantná alebo približne konštantná volí sa λ=1
2) Ak hodnota r rastie lineárne s m volí sa $\lambda = 0$
3) Hodnota r rastie rýchlejšie ako lineárne, $\lambda < 0$
4) Ak r pomalšie ako lineárne, volí sa $0 < \lambda < 1$

Procesy ARIMA(p, d, q) sa často označujú ako integrované procesy rádu d s označením I(d). V jednoduchosti môžeme povedať, že ak po diferencovaní časového modelu, pomocou transformácie, aplikujeme model ARMA(p, q). Vo vzťahu k pôvodnému časovému radu sme vytvorili ARIMA(p, d, q).

časový rad | diferenciácia | ARMA(p, q) = časový rad | ARIMA(p , d, q)

Matematické vyjadrenie konkrétneho modelu

Príkladom jednoduchého modelu ARIMA je ARIMA(0,1,0), ktorý vznikne diferencovaním pôvodného časového radu. Zo zápisu (0, 1, 0) sú zrejmé hodnoty jednotlivých rádov: p = 0, d = 1, q = 0. Dosadením hodnôt parametrov p, d, q do (2.28) a využitím vzťahu (2.29) môžeme model pomocou operátora spätného posunutia vyjadriť v tvare:

$$\phi_{p=0}(\mathrm{B})(1 - \mathrm{B})^{d=1}\mathrm{y}_t = \mathrm{c} + \theta_{q=0}(\mathrm{B})\mathrm{a}_t . \tag{2.31}$$

Keďže rády p a q sú rovné nule, model nebude obsahovať autoregresný operátor $\phi(\mathrm{B})$ a ani operátor kĺzavých súčtov $\theta(\mathrm{B})$, zápis sa nám zjednoduší na tvar:

$$(1 - \mathrm{B})^{d=1}\mathrm{y}_t = \mathrm{c} + \mathrm{a}_t , \tag{2.32}$$

kde c predstavuje konštantu rovnú strednej hodnote μ časového radu, a ktorej významnosť by sme mali pred zápisom do vyjadrenia modelu otestovať. Zápis modelu ARIMA(0, 1, 0) so štatisticky významnou konštantou c (2.32) môžeme napokon vyjadriť bez operátora spätného posunutia takto:

$$y_t = c + y_{t-1} + a_t . \tag{2.33}$$

Model (2.33) označujeme, ako model s tvarom náhodnej prechádzky.

2.4 Nelineárny proces a nelineárne modely

Nelineárny proces sa dá popísať nasledujúcou štruktúrou [11]:

$$y_t = g\left(a_{t-1}, a_{t-2}, ...\right) + a_t h\left(a_{t-1}, a_{t-2}, ...\right) \tag{2.34}$$

kde **stredná podmienená hodnota** tohto procesu je

$$E\left[y_t \mid y_{t-1}\right] = g\left(a_{t-1}, a_{t-2}, ...\right) \tag{2.35}$$

a **podmienený rozptyl** (variancia)

$$\text{var}[y_t \mid y_{t-1}] = E\left[\left\{ \left(y_t - E[y_t]\right) \mid y_{t-1} \right\}^2\right] = \left\{h\left(a_{t-1}, a_{t-2}, ...\right)\right\}^2 \tag{2.36}$$

Funkcia $g(.)$ zodpovedá podmienenej strednej hodnote procesu y_t , a funkcia $h(.)$ je koeficientom úmernosti medzi zmenou y_t a a_t [11]. Zo vzťahu (2.35) vyplýva, že ak je $g(.)$ nelineárna tak potom je model nelineárny v strednej hodnote. A podľa vzťahu (2.36) ak je nelineárna funkcia $h(.)$ potom je aj model nelineárny v rozptyle.

Podmienený rozptyl (variancia)

Pri určení podmieneného rozptylu vychádzame z teórie pravdepodobnosti, konkrétne podmienenej pravdepodobnosti [8], kde náhodné premenné sú na sebe závislé. Majme dve závislé náhodné premenné X a Y. Potom pre rozptyl premennej X podmienenej Y platí:

$$\text{var}(X \mid Y) = E[X - E[(X \mid Y)] \mid Y] \tag{2.37}$$

V prípade, ak by X a Y boli nezávislé: $\text{var}(X \mid Y) = \text{var}(X)$.

Predstavme si teraz nasledujúci príklad modelu nelineárneho v strednej hodnote avšak lineárneho v rozptyle. je NMA model (Nonlinear Moving Average), ktorý popisuje vzťah

$$y_t = a_t + \alpha a_{t-1}^2 \tag{2.38}$$

kde $g(.) = \alpha a_{t-1}^2$ a $h(.) = 1$ [11].

2.4.1 Heteroskedasticita

Klasický lineárny regresný model je charakterizovaný požiadavkou konečného a konštantného rozptylu náhodných zložiek $\mathrm{D}\left(y_t\right) = \sigma^2$ pre $\forall$ t. Taktiež táto požiadavka platí pre rezíduá modelu, $\mathrm{D}\left(a_t\right) = \sigma^2$ pre $\forall$ t, odborne nazýva homoskedasticita. Proces s konečným a konštantným rozptylom zložiek môžeme označiť ako homoskedastický.

Naopak, v prípade náhodného výberu pozorovaní, v ktorom dochádza k veľkým zmenám hodnôt (vysoká variabilita) hovoríme o opačnom jave heteroskedasticite [17]. V takomto prípade je porušená podmienka a môžeme zapísať $\mathrm{D}\left(a_t\right) \neq \sigma^2$ pre $\forall$ t.

Za príčiny výskytu heteroskedasticity môžeme označiť samotné pozorovania (v našom prípade sledovaný časový rad). Významnú úlohu hrá aj existencia chýb meraní. Ďalšou častou príčinou je nesprávna špecifikácia modelu, spočívajúcu vo vynechaní významnej premennej, alebo v použití skupinových priemerov namiesto pôvodných pozorovaní [17]. Vo väčšine prípadov je však ťažké určiť príčinu heteroskedasticity, preto sa ju snažíme odstrániť rôznymi transformáciami (napr. logaritmická). V prípade výskytu heteroskedasticity v samotnej podstate dát, je vhodnejšie pri procese vytvárania modelu zohľadniť túto vlastnosť a využiť modely schopné zachytiť heteroskedasticitu.

Heteroskedasticita je, stručne, pojem pre v čase premenný a nekonečný rozptyl náhodných porúch a reziduí.

2.4.1 Proces ARCH(*q*)

V posledných rokoch vývoj rôznych odvetví, v ktorých sa aplikuje analýza časovej rady podnecuje k využívaniu nelineárnych modelov. Predpokladá sa, že aj zdanlivá nezávislosť volatívnych javov (nestálych alebo s veľkou variabilitou) môže byť pomocou analýzy časových radov predikovaná na základe existujúcej nelineárnej závislosti. Ukázalo sa, že ARCH modely sú vhodným nástrojom pre takúto analýzu. Skratka ARCH pochádza z anglického názvu pre autoregresívnu podmienenú heteroskedasticitu (Autoregressive Conditional Heteroskedasticity). Zaviedol ich v roku 1982 Engle [11].

Ak si všimneme napríklad vzťah (2.24) pre lineárny proces AR(*p*), reziduálna zložka a_t je aditívne pridaná k inováciám. Keďže a_t má konštantný a konečný rozptyl (2.22), AR(*p*) model je schopný úspešne modelovať proces iba pod podmienkou jeho konštantného rozptylu.

ARCH model však dovoľuje modelovať zmenu podmieneného rozptylu a zmeny minulých odchýlok v čase. Okrem toho rozlišuje medzi podmieneným rozptylom a nepodmieneným rozptyl. Nepodmienený rozptyl na rozdiel od podmieneného zachováva konštantný [16]. Tieto vlastnosti umožňujú čiastočne popísať nelinearitu, poradiť si s hromadením chýb a predovšetkým predikovať náhle zmeny [11].

Podľa Engleho, ARCH proces matematicky definuje vzťah:

$$y_t = a_t\sqrt{\alpha a_{t-1}^2}\,, \tag{2.39}$$

kde funkcia $g(.) = 0$, a $h(.) = \sqrt{\alpha a_{t-1}^2}$, preto je tento model (na rozdiel od NMA modelu) nelineárny v rozptyle, ale lineárny v strednej hodnote [11].

Na základe vyššie definovaného procesu (2.39) navrhol Engle nasledujúci model, ktorý má za úlohu zachytiť sériovú koreláciu javov s vysokou variabilitou [11]:

$$\sigma^2 = \omega + \alpha(L)\eta_t^2\,, \tag{2.40}$$

kde $\alpha(L)$ označuje polynóm reprezentujúci oneskorené parametre, $\eta_t \mid \Psi_{t-1}$ sú inovácie s rozdelením $\sim N(0,\sigma_{t-1}^2)$. ARCH model charakterizuje distribúciu stochastických odchýlok a_t podmienených skutočnými hodnotami série premenných $\Psi_{t-1} = \{y_{t-1}, x_{t-1}, y_{t-2}, x_{t-2}, ...\}$.

Po úpravách v (2.40) môžeme zapísať ekvivalentnú reprezentáciu ARCH (*q*) model takto:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 \tag{2.41}$$

V tomto prípade pre nesystematické zložky a_t pre $\forall$ t platí $a_t = \eta_t\sqrt{\sigma_t^2}$, t.j. majú premenlivý (nekonštantný) rozptyl. Označenie η_t je pre inovácie, ktoré majú charakter šumu $\eta_t = N(0,\sigma_t^2) \sim N(0,1)$ [19], kde rozloženie $N(0,1)$ indikuje Gaussovský šum. Zo vzťahu (2.41) je vidieť časovú závislosť rozptylu.

Podľa literatúry [7, 11], je možné bez straty všeobecnej platnosti definovať jednoduchý ARCH model:

$$y_t = a_t a_{t-1}^2 \tag{2.42}$$

Pripomenieme si, že a_t predstavuje biely šum s vlastnosťami popísanými v podkapitole 2.2.1.

Čo sa týka vlastností procesu (2.42), popíšeme si ich v nasledujúcich výpočtoch [7]:

$$E(y_t) = E\left(a_{t-1}^2 a_t\right) = E\left(a_{t-1}^2\right) E(a_t) = 0 \tag{2.43}$$

$$\text{var}(y_t) = E\left(y_t^2\right) = E\left(a_{t-1}^4 a_t^2\right) = E\left(a_{t-1}^4\right)^* E\left(a_t^2\right) = \left(3\sigma_a^4\right)\sigma_a^2 = 3\sigma_a^6 \tag{2.44}$$

$$E(y_t \mid y_{t-h}) = E\left(a_{t-1}^2 a_t a_{t-h-1}^2 a_{t-1}\right) = E(a_t)\, E\left(a_{t-1}^2 a_{t-h-1}^2 a_{t-h}\right) = 0, \text{ pre } h > 0 \tag{2.45}$$

* využili sme poznatok o normovanom momente 4. rádu $E\left(a_t^4\right)$ nazývaného koeficient špicatosti [6], u ktorého pri normálnom rozdelení a_t platí (2.46).

$$m_4^*(a_t) = E\left(\frac{a_t^4}{\sigma_a^4}\right) = \frac{E\left(a_t^4\right)}{\sigma_a^4} = 3 \;\rightarrow\; E\left(a_t^4\right) = 3\sigma_a^4 \tag{2.46}$$

Z uvedených vzťahov (2.43),(2.44),(2.45) sú zrejmé vlastnosti procesu y_t:

a) kovariančne (slabo) stacionárny
b) y_t má nepodmienený rozptyl var(y_t) = $3\sigma_a^2$ (nezávisí na čase)
c) y_t má podmienený rozptyl hodnotou y_{t-1}

$$\text{var}(y_t \mid y_{t-1}) = \text{var}\left(a_{t-1}^2 a_t \mid y_{t-1}\right) = \sigma_a^2 a_{t-1}^4 \tag{2.47}$$

na čase závisí prostredníctvom faktora a_{t-1}^4. Vzťah (2.47) dokazuje tvrdenie o nelinearite ARCH modelu v rozptyle, ktoré bolo spomenuté pri definícii.

2.4.2 Proces GARCH (*p*, *q*)

Veľkou nevýhodou ARCH(q) modelu je, že rastúcim rádom p vzrastá náročnosť výpočtu, preto sa v praxi častejšie používa generalizovaný autoregresný podmienený heteroskedastický model GARCH (p, q) [11]:

$$\sigma_t^2 = \omega + \alpha(L)\eta_t^2 + \beta(L)\sigma_{t-1}^2 . \tag{2.48}$$

Ak zvážime podobnosť štruktúry s modelom ARMA(p, q), tak v modeli GARCH (p, q) polynóm $\alpha(L)$ rádu "q" zodpovedá kĺzavým súčtom MA(q) a polynóm $\beta(L)$ rádu "p" korešponduje s prvkom autoregresný proces AR(p). Lepšie pochopenie vzťahu (2.48) poskytuje obdobný zápis (2.50).

Bollerslev vo svojej práci [16] vysvetľuje GARCH(p, q) proces nasledovne: nech a_t označuje v čase diskrétny stochastický proces s reálnymi hodnotami a Ψ_t označuje informačný súbor (sériu odchýlok oboru σ) v čase t. GARCH(p, q) proces je potom daný nasledujúcimi vzťahmi:

$$a_t \mid \Psi_{t-1} \sim N(0, \sigma_t^2), \tag{2.49}$$

$$\begin{aligned} \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i}^2 + \sum_{j=1}^{p} \beta_i \sigma_{t-i}^2 = \\ &= \alpha_0 + \alpha(L) a_t^2 + \beta(L)\sigma_t^2 \end{aligned} \tag{2.50}$$

kde

$$\begin{aligned} &p \geq 0, \qquad q > 0 \\ &\alpha_0 > 0, \qquad \alpha_i \geq 0, \qquad i = 1,...,q, \\ &\beta_i \geq 0, \qquad i = 1,...,p. \end{aligned}$$

Ak p = 0, proces sa redukuje na ARCH(q) proces. V prípade $p = q = 0$ nám zostane iba jednoduchý biely šum (zložka a_t). Zatiaľ, čo v ARCH(q) procese je podmienený rozptyl špecifikovaný len ako lineárna funkcia minulých vzoriek podmieneného rozptylu,

GARCH(p, q) proces zavádza aj vplyv oneskorených podmienených rozptylov. Táto vlastnosť sa dá vnímať ako druh adaptívne učiaceho sa mechanizmu.

Proces definovaný v (2.50) je stacionárny s vlastnosťami [1]:
$E(a_t) = 0$, $\text{var}(a_t) = \alpha_0(1 - A(1) - B(1))^{-1}$ a $\text{cov}(a_t, a_s) = 0$ pre $t \neq s$ v jedinom prípade ak $A(1) + B(1) < 1$.

Najjednoduchší ale v praxi najvyužívanejší GARCH model je bezpodmienečne GARCH(1,1) [11, 16, 18]. Matematický zápis tohto modelu dostaneme odvodením od všeobecného vzorca (2.50):

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{2.51}$$

pričom platí $\alpha_0 > 0$, $\alpha_1 \geq 0$ a $\beta_1 \geq 0$.

3 Vytváranie modelu v Box-Jenkinsovej metodológií

Vlastná konštrukcia modelu v tejto metodológií sa dá rozdeliť na tri etapy: **identifikácia**, **odhad parametrov** a **diagnostická kontrola** (testovanie, verifikácia modelu). Postupy, popísané v nasledujúcich riadkoch demonštruje praktický príklad uvedený v prílohe C.

3.1 Identifikácia modelu

Identifikácia modelu je jedna z najťažších úloh pri výstavbe modelov. Podstatou identifikácie je výber vhodného modelu a určenie jeho rádu. Celkový proces identifikácie závisí na rôznych faktoroch pojatých v úvodnej kapitole.

V prvom rade, je vhodné preskúmať graf časovej rady. V mnohých prípadoch je možné na prvý pohľad rozpoznať výskyt jednotlivých zložiek, napríklad trendu. Hoci ide o dosť subjektívny prístup, na jeho základe sa dajú vylúčiť, či prijať možnosti, nutnosť, úprav. V prípade výskytu trendu, je zrejmé, že sledovaná rada nie je stacionárna, je nutné ju stacionarizovať pomocou diferenciácie. Prípadne sa môžu previesť iné úpravy ako linearizácia alebo Box-Jenkinsove transformácie. Transformácie je vhodné vykonať pred vlastným diferencovaním. pretože pri diferencovaní môžeme dostať záporné hodnoty [6].

Ak si nie sme istý či je časový rad stacionárny, a je treba ho stacionarizovať, zistíme si to pomocou výpočtu jeho odhadov ACF a PACF. V prípade, že hodnoty ACF a PACF sú v prvom oneskorení $k = 1$ veľmi blízke hodnote 1 a zároveň ostatné hodnoty klesajú veľmi pomaly, analyzovaný časový rad je nestacionárny [6] a musíme ho stacionarizovať.

V prípade, že stredná hodnota stacionarizovaného časového radu je nenulová, musí sa tento proces vycentrovať. Princíp spočíva v nahradení radu y_t radou $y_t - \overline{y}_t$. Ak existujú pochybnosti o nenulovosti strednej hodnoty, používajú sa bežné štatistické testy. Napríklad porovnanie aritmetického priemeru $\overline{y}_t$ s dvojnásobkom smerodajnej odchýlky tohto priemeru (viď [10]).

Ďalším krokom je odhad ACF a PACF. Vo väčšine prípadov stačí niekoľko prvých hodnôt týchto funkcií. Na základe, ktorých sa určí identifikačný bod k_0 (kapitola 2 funkcie ACF, PACF) , ktorý však existovať vôbec nemusí. Hodnoty rádov p a q procesov AR a MA korešpondujú s identifikačným bodom. Nakoľko odhady ACF a PACF môžu byť navzájom silne korelované, nie je dobré spoliehať sa na prvý stanovený model ale vyskúšať viacero modelov.

Dobrý popis a pripomienky k postupu pri konštrukcii ARIMA modelov nájdeme napríklad v literatúre [14].

3.2 Odhad parametrov modelu

Po identifikácií modelu sa odhady jeho parametrov postupne upresňujú pomocou iteračných postupov. Najčastejšie sa používa metóda najmenších nelineárnych štvorcov.

Predpokladáme , že pre danú časovú radu bol zvolený model ARMA(1,1), t.j. model s dvoma parametrami v tvare $y_t = \phi_1 y_{t-1} + a_t + \theta_1 a_{t-1}$. Hľadáme také odhady parametrov ϕ_1 a θ_1, pre ktoré nadobúda funkcia (3.1) minimum.

$$S(\phi_1, \theta_1) = \sum_{t=1}^{n} a_t^2 (\phi_1, \theta_1) \tag{3.1}$$

Hľadanie minima tejto funkcie je prácne a je záležitosťou numerickej matematiky. Z toho dôvodu sa v praxi spolieha na vhodný štatistický software, ktorý je schopný tieto

parametre vypočítať. Po výpočte odhadu parametrov sa určuje ich presnosť, ktorú by mal dobrý štatistický software zvládnuť.

3.3 Diagnostická kontrola (overenie) modelu

Cieľom tejto etapy je testovanie na základe, ktorej sa potvrdí vhodnosť a správnosť navrhnutého modelu. Metód na testovanie existuje niekoľko. Toto sú najpoužívanejšie:

1) Metóda **preparametrizovania** modelu. Ak sú smerodajné odchýlky odhadnutých parametrov, prípadne nesystematickej zložky a_t príliš vysoké, odporúča sa použiť nový model s väčším počtom parametrov. V prípade ak sa odhady výrazne nelíšia od nuly, je treba previesť redukciu počtu parametrov.
2) Metóda **odhadnutých reziduí.** Princíp bude demonštrovaný na modeli ARMA(1,1). Ak $\hat{\phi}_1$ a $\hat{\theta}_1$ sú odhady parametrov modelu, potom odhadnuté rezíduá sú veličiny $\hat{a}_t = y_t - \hat{\phi}_1 y_{t-1} - \hat{\theta}_1 y_{t-1}$ .Pre časový rad týchto reziduí vypočítame odhady reziduálnej autokorelačnej funkcií pomocou vzťahu

$$r_k = \frac{\sum_t \hat{a}_t \hat{a}_{t-k}}{\sum_t \hat{a}_t^2} \text{ [6].} \tag{3.2}$$

Následne musíme otestovať, či niektorý z autokorelačných koeficientov neprekračuje medze 95% intervalu spoľahlivosti. To vykonáme pomocou porovnania hodnôt $r_k(\hat{a})$ s dvojnásobkami ich smerodajných odchýlok $2\sigma\left(r_k\left(\hat{a}\right)\right)$. Ak pre niektorú hodnotu k platí vzťah

$$r_k\left(\hat{a}\right) > 2\sigma\left(r_k\left(\hat{a}\right)\right) \tag{3.3}$$

potom $r_k(\hat{a})$ prekračuje medze 95% intervalu spoľahlivosti, nesystematická zložka $\hat{a}$ je korelovaná a overovaný model je neadekvátny.

3) Metóda založená na portmanteau test. Ide o testovanie štatistickej hypotézy adekvátnosti navrhnutého modelu. Používa sa štatistické kritérium v tvare

$$Q = n\sum_{k=1}^{K} r_k^2\left(\hat{a}\right) \tag{3.4}$$

kde n je dĺžka časovej rady a K je prirodzené číslo porovnateľné s hodnotou $\sqrt{n}$. Pre model ARMA (p, q) má portmanteau štatistika Q asymptotické rozdelenie $|\chi^2_{K-p-q}$. Ak pre danú hladinu významnosti α a experimentálne určené Q_{exp} platí $Q_{\text{exp}} > \chi^2_{K-p-q}(\alpha)$ [7], overovaný model je neadekvátny na hladine významnosti α. Existuje však množstvo iných štatistík na určenie adekvátnosti modelu

4 Základné modely sieťovej prevádzky

4.1 Charakteristika sieťových modelov

Základnými parametrami, ktoré charakterizujú prevádzku v dátových sieťach, sú: rozdelenie dĺžky (veľkosti) paketov a rozdelenie intervalov príchodu, t.j. časových úsekov medzi dvomi po sebe doručenými paketmi *n-1* a *n* [1]. Existuje, samozrejme, mnoho ďalších parametrov, tie však nie sú natoľko dôležité pre stanovenie modelov, prípadne úzko súvisia so špecifickou topológiou siete. Čo sa týka rozdelenia veľkosti paketov, bolo definovaných už množstvo jednoduchých a osvedčených modelov popisujúcich tento problém. Oveľa zložitejším problémom je vytvoriť efektívny model, ktorý popíše distribúciu intervalov príchodu.

Staršie sieťové modely sú založené na jednoduchých rozdeleniach pravdepodobnosti funkčných za predpokladu, že v zloženej sieťovej prevádzke majú tzv. zhluky tendenciu vymiznúť a jej celkový priebeh sa vyhladí. To znamená, že s rastúcim počtom zdrojov klesá výskyt zhlukov. Niektoré modely zhluky v sieti prakticky vôbec neberú do úvahy (Poissonov model) [1]. Ako sa časom a nástupom dátových sietí ukázalo, takéto modely nie sú príliš vhodné vzhľadom na fakt, že skutočná sieťová prevádzka vykazuje výskyt zhlukov a to v rôznych časových rozsahoch [2].

4.2 Poissonov model

Poissonov model bol zavedený ešte v ére telefónnych ústrední. Z analytického hľadiska ide o atraktívny model, vzhľadom na to, že nemá žiadne pamäťové nároky [2] (počet príchodov v časovom úseku nasledujúcom po *t* nie je závislý na počte príchodov pred *t*). Pomocou správnej voľby jediného parametru λ, sa dá aplikovať na prevádzku väčšiny dátových sietí, avšak s obmedzením na časový úsek, ktorý má relatívne hladký priebeh a nevykazuje zhlukov osť [1].

Parameter λ popisuje intenzitu príchodu požiadaviek, dlhodobý priemer počtu prichádzajúcich požiadaviek a má rozmer čas^{-1}. Ako už bolo spomenuté, predpokladá sa, že jednotlivé príchody paketov sú nezávislé. Intervaly medzi príchodmi majú exponenciálne rozdelenie so strednou hodnotou $\frac{1}{\lambda}$. To znamená, že počet príchodov paketov za časový úsek $t = t2 - t1$ má Poissonovské rozdelenie so strednou hodnotou λt [1].

Združením viacerých nezávislých Poissonovských procesov dostaneme nový Poissonov proces s intenzitou rovnou súčtu čiastkových intenzít $\lambda' = \sum\lambda$ [1]. Pre exponenciálne rozdelenie platí, že aj rozptyl je rovný λ.

Jednoduchým spôsobom, ako zistiť, že určitý prúd dát v sieti má Poissonovské rozdelenie, je zobrazenie histogramu intervalov príchodu a overenie, že sa jedná o exponenciálne klesajúcu funkciu. Vyplýva to zo vzťahu pre Poissonovo rozdelenie (1.1) [3], ktorý popisuje pravdepodobnosť príchodu *n* paketov v čase *t*.

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{4.1}$$

Člen $e^{-\lambda t}$ vo vzťahu (1) reprezentuje klesajúcu exponenciálnu funkciu. Analogicky $e^{-\lambda t}$ definuje pravdepodobnosť príchodu 0 paketov v čase *t*, po dosadení $n = 0$ do (1).

4.3 Poissonov zložený model

Od jednoduchého Poissonovho modelu, ktorý popisuje rozdelenie intervalov príchodu v časovej mierke, sa zložený model odlišuje v rozšírení o priestorový popis toku

dát. To znamená, že popisuje príchod série alebo zhlukov paketov, ktorá dorazila v jednom časovom okamihu [2]. Prípadne môže sériu paketov reprezentovať jeden paket s definovanou veľkosťou (jeho dátový objem). Zatiaľ čo, intervaly medzi príchodmi sérií majú exponenciálne rozdelenie. Veľkosť doručených paketov má rozdelenie geometrické.

Z matematického hľadiska je zrejmé, že na výpočet zloženého Poissonovho modelu je potrebné použiť dva procesy. Sú teda potrebné dva parametre, λ intenzita príchodu a δ parameter z intervalu (0,1), charakterizujúci dávku paketov. Priemerný počet paketov v jednej sérii (dávke) je jeho prevrátená hodnota $\frac{1}{\delta}$.

4.4 Obmedzenie Poissonovských modelov

Vyššie diskutované rozšírenie Poissonovho modelu je jednou z ciest ako modelovať zhlukovosť, vyskytujúcu sa v sieťovej prevádzke. Pre Poissonovské modely [1, 3] však platí centrálna limitná veta [1, wiki]. Tento teorém vysvetľuje fakt, že pri zoskupení veľkého množstva dát s istým rozdelením alebo dát s rôznymi rozdeleniami sa analyzované rozdelenie priblíži normálnemu rozdeleniu. S podmienkou, že hodnoty (analyzované dáta) sú nezávislé a majú nekonečný rozptyl. Ak sa teda dostatočne zvýši miera agregácie, zhluky sa priemerovaním nad dlhým časový rozsahom vyhladia a napokon (v grafickom zobrazení) dosiahnu tvar plochej krivky v strednej hodnote rozdelenia. Korelácia pozorovaných hodnôt sa v takomto prípade stane bezvýznamnou. To však neplatí v reálnych sieťach, kde sa dátová prevádzka vyznačuje sebe–podobnosťou, ktorá v niektorých prípadoch vykazuje fenomén dlhodobej závislosti (Long Range Dependency - LRD).

5 Sebe–podobnosť v celosvetovej sieťovej prevádzke

Na rozdiel od modelov spomenutých v Kapitole 4. skutočná sieťová prevádzka vykazuje variabilitu vo všetkých časových mierkach. To znamená, že zhlukovosť môžeme pozorovať aj vo väčšom časovom úseku alebo v celom rozsahu. Štatisticky sa dá takáto prevádzka popísať pomocou pojmu sebe–podobnosť. Sebe–podobnosť je vlastnosť súvisiaca s fraktálom - objekt, ktorého vzhľad zostáva vždy rovnaký bez ohľadu na mierku pod akou ho pozorujeme [4].

5.1 Definícia sebe–podobnosti

Ak $X = (X_t ;\ t = 0,1,2, \ldots)$ je kovariančne stacionárny stochastický proces [3] so strednou hodnotou μ, rozptylom σ^2 a autokorelačnou funkciou $r(k)$, $k = 0,1,2, \ldots$, potom $X_k^{(m)}$, pre všetky $m \in N$, označuje nový časový rad získaný spriemerovaním originálnych časových radov X (2.1).

$$X_k^{(m)} = \frac{1}{m}\left(X_{km-m+1} + \ldots + X_{km}\right), \quad k \in N \tag{5.1}$$

$X^{(m)}$ definuje kovariančne stacionárny proces m agregovaných časových radov s odpovedajúcou autokorelačnou funkciou $r^{(m)}$.

Predpokladáme, že autokorelačná funkcia procesu X má tvar:

$$r(k) \sim k^{-\beta} L_1(k), \quad k \to \infty, \tag{5.2}$$

kde parameter $0 < \beta < 1$ a pre funkciu L_1 platí $\lim_{t\to\infty} \frac{L_1(tx)}{L_1(t)} = 1$ pre všetky x > 0. Príkladmi takej funkcie sú $L_1(t) = konšt.$ alebo $L_1(t) = \log(t)$.

Proces X sa nazýva **úplne sebe–podobný** s parametrom sebe–podobnosti

$$H = 1 - \beta/2, \tag{5.3}$$

ak pre všetky $m \in N$ platí, že agregovaný proces $X^{(m)}$ má po stanovení novej mierky pomocou m^H rovnaké rozdelenie ako X. Podmienku sebe–podobnosti popisuje vzťah (2.4), v ktorom $=^d$ označuje rovnosť distribúcií.

$$X_t =^d \frac{1}{m^H}\left(X_{km-m+1} + \ldots + X_{km}\right), \quad \forall m \in N \tag{5.4}$$

Proces je úplne **sebe–podobný druhého rádu** s parametrom H (2.3), ak $\forall m \in N$ má $\frac{1}{m^H}\left(X_{km-m+1} + \ldots + X_{km}\right)$ rovnaký rozptyl a autokoreláciu ako proces X [5]. To znamená, že $\forall m \in N$, $\mathrm{var}(X^{(m)}) = \sigma^2 m^{-\beta}$ a zároveň $r^{(m)}(k) = r(k) = \frac{1}{2}\sigma^2\left(|k|^{2-\beta}\right)$, $k = 0,1,2,\ldots$. Kde $\sigma^2(f)$ je druhý centrálny diferenčný operátor použitý na funkciu $\sigma^2(f(k)) = f(k+1) - 2f(k) + (k-1)$.

Asymptoticky sebe–podobný druhého rádu s parametrom H (2.3) sa nazýva proces, ak sa pre veľké hodnoty m, autokorelácia $r^{(m)}$ asymptoticky zhoduje s autokoreláciou r procesu X. Matematická definícia: $r^{(m)}(k) \to \frac{1}{2}\sigma^2\left(|k|^{2-\beta}\right)$ pre $m \to \infty$, $k = 0,1,2,\ldots$. Príkladmi asymptoticky sebe–podobných procesov druhého rádu sú frakčné autoregresívne procesy s integrovaným kĺzavým priemerom alebo v skratke frakčné ARIMA(p, d, q) s parametrom sebe–podobnosti $H = d + 1/2$, kde $0 < d < 1/2$.

Intuitívne najvýznamnejšou vlastnosťou sebe–podobných procesov je, že so zvyšujúcou sa agregáciou korelácia nemá tendenciu sa degenerovať. Naopak

u konvenčných stochastických modelov [1, 2] sa proces $X^{(m)}$, so zvyšujúcou sa mierou agregácie ($m \to \infty$), mení na šum: $r^{(m)}(k) \to 0, \quad k = 1,2,3...$.

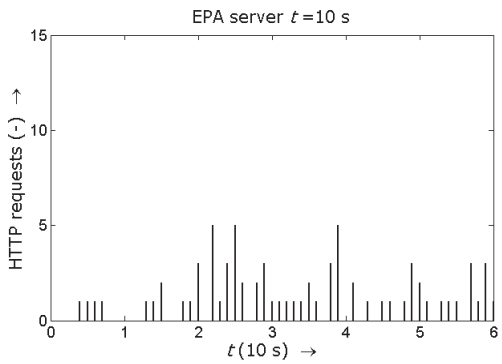
Vyššie uvedené definície sebe–podobnosti vychádzajú z matematicky pohodlnejšieho vyjadrenia (2.5) podmienky sebe-podobnosti spojitého stochastického procesu $X = (X_t ; t \geq 0)$ s nulovou strednou hodnotou a stacionárnym prírastkom, $\forall\ a > 0$,

$$X_{at} =^{d} a^{H} X_t , \tag{5.5}$$

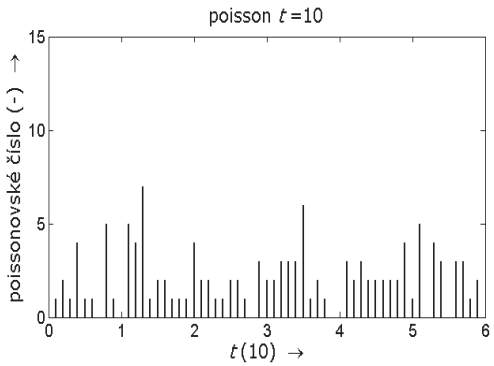
kde obdobne $=^{d}$ označuje rovnosť distribúcií.

5.2 Vizuálna demonštrácia sebe-podobnosti v sieťovej prevádzke

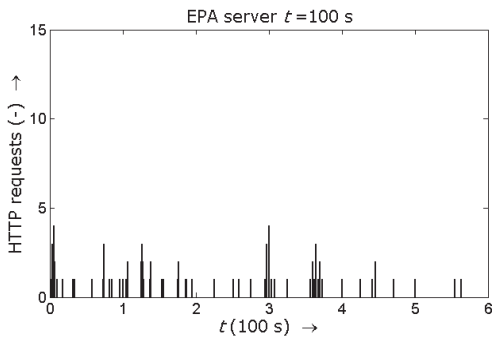
Vyššie popísaná teória sebe-podobnosti, sa dá vizuálne objasniť na praktických ukážkach. Porovnaním grafov zachytenej sieťovej prevádzky so synteticky generovanou prevádzkou. Nasledujúci súbor grafov (Obr.č.1) okrem vizualizácie sebe-podobnosti dokazuje jej existenciu v sieti Internet a poukazuje na zlyhanie Poissonovského procesu zachytiť vlastnosť sebe-podobnosti.



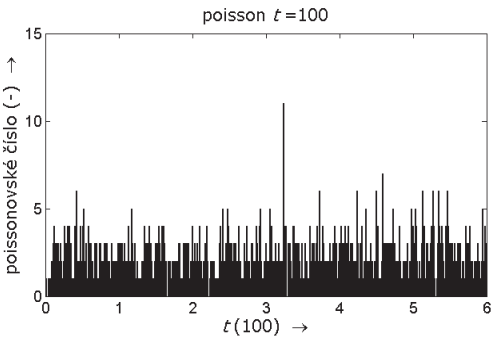
1(a)



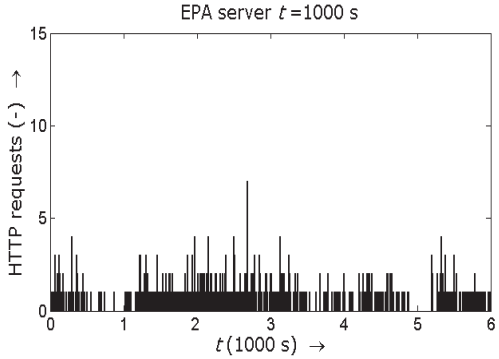
2(a)



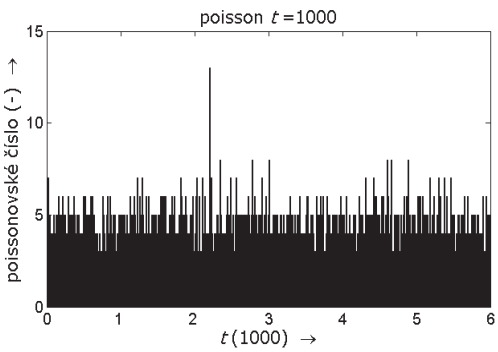
1(b)



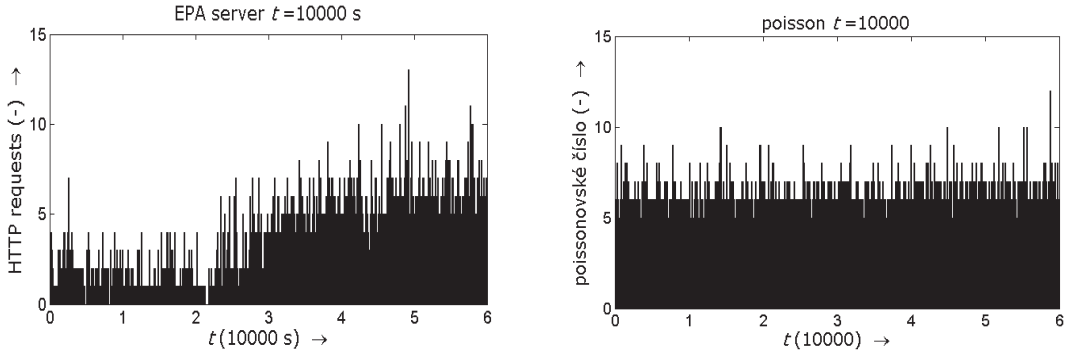
2(b)



1(c)



2(c)



1(d) 2(d)

Obr. č. 1*: Súbor grafov demonštrujúcich sebe-podobosť. V ľavom stĺpci, (1(a), 1(b), 1(c), 1(d)), časového radu HTTP požiadaviek servera EPA. V pravom stĺpci (2(a), 2(b), 2(c), 2(d)) náhodne generované čísla s Poissonovským rozdelením.*

Budeme si všímať predovšetkým rozdelenie dvoch rozličných procesov a pri zmene mierky zobrazenia. Na ľavej strane (1(a), 1(b), 1(c), 1(d)) je graficky znázornený intervalový časový rad s intervalom pozorovania 1s, získaný z verejne dostupného archívu Internetovej prevádzky ITA („http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html"). Jedná sa o pozorovanie prichádzajúcich HTTP požiadaviek na jeden zo serverov EPA (Environmental Protection Agency). Na pravej strane sú do grafov (2(a), 2(b), 2(c), 2(d)) vynesené hodnoty náhodných čísiel s Poissonovským rozdelením a parametrom $\lambda = 2$. Súbor týchto hodnôt bol generovaný v prostredí MATLAB pomocou funkcie ***poissrnd***. Popis tejto funkcie je dostupný v dokumentácii na webovej stránke spoločnosti Mathworks („http://www.mathworks.com/help/toolbox/stats/poissrnd.html"). Nakoľko Poissonovské náhodné čísla neboli generované v reálnom čase, neuvádzam pri veličine t jednotku, ale iba jej rozmer . Môžeme si ju však predstaviť ako jednotku času 1s.

Vo vertikálnom smere sa grafy líšia v časovej mierke, ktorá sa smerom od indexov a–d zväčšuje. Tieto dva 1(.) a 2(.), môžeme napísať procesy, sa pri malej mierke (t=10) prakticky nedajú od seba rozoznať. So zväčšujúcou sa mierkou si môžeme všimnúť, že jeho rozdelenie procesu 2(.) sa približuje normálnemu rozdeleniu, t.j. má charakter bieleho šumu. Tento fakt popisuje už spomínaná centrálna limitná veta a pri modelovaní prevádzky s vysokou variabilitou je nežiaduci. Naopak v procese 1(.) je aj po zväčšení mierky (agregácií hodnôt) pozorovateľná určitá variabilita.

Pre úplnosť uvádzam, že adekvátnejšie porovnanie reálnej a synteticky generovanej prevádzky je popísané v literatúre [5] Adekvátnejšie z dôvodu grafického zobrazenia, kde je [5] vynesený časový rad s rôzne veľkým intervalom pozorovania, to lepšie vystihuje agregáciu. V našom prípade (Obr.č.1) sa mení iba mierka zobrazenia (horizontálna os) nie interval pozorovania. Hodnoty na vertikálnej osi v každom grafe stále odpovedajú počtu paketov za 1s So zmenou intervalu pozorovania, napríklad z 1s na 10s, by boli na vertikálnu os vynesené podstatne vyššie hodnoty rovnajúce sa sume zachytených paketov počas 10 s Na vysvetlenie charakteru reálnej prevádzky a dôkaz sebe-podobnosti je Obr.č.1 postačujúci.

5.3 Dlhodobá závislosť (LRD)

Sebe–podobný proces vykazuje dlhodobú závislosť. To znamená, že autokorelačná funkcia takéhoto procesu má charakter pomaly klesajúcej funkcie. Klesá pomalšie ako exponenciálne a krivka jej rozpadu je skôr mocninová. V praxi to znamená,

že stavy (udalosti) v systéme procesu, ktoré nastali pred časovým okamihom t , budú mať vplyv na proces ešte ďaleko v budúcnosti za okamihom t.

5.4 Vplyv riadenia a kontroly TCP prevádzky

Po objavení sebe–podobnosti v sieťovej prevádzke, boli publikované výsledky niekoľkých štúdií demonštrujúcich podcenenie variability (zhlukovosti) v TCP prevádzke [1, 2, 4, 5, 13]. Dramaticky zmenil situáciu vývoj systémov určených na riadenie a dohľad zahltenia v dátových sieťach najmä čo sa týka TCP prevádzky. Toky paketov už viac nevykazujú plnú nezávislosť, čo bol dôležitý predpoklad pre tradičné sieťové modely [1]. Prevádzka je podmienená rôznymi stavmi v sieti, ktoré fungujú ako spúšťače. Napríklad zahltenie siete, ktoré aktivuje systémy pre riadenie zahltenia. Táto závislosť spôsobuje obmedzenia vo využití niektorých modelov a mala by byť rešpektovaná.

Pozor, je dôležité uvedomiť si, že riadenie a kontrola TCP prevádzky nie je príčinou výskytu sebe–podobnosti, ani ju neodstráni a ani jej vplyv nezmierni [1].

V oblasti komunikácií, má a bude mať manažment a optimalizácia prevádzky v sieťach veľkú úlohu. Nevyhnutným predpokladom pre efektívne projektovanie, riadenie, a dohľad siete je predovšetkým ich analýza a predikcia, ako výsledok analýzy. Pre tieto účely boli zavedené **modely časových radov,** ktoré sa vyznačujú schopnosťou zachytiť špecifické vlastnosti sieťovej prevádzky. V nasledujúcej kapitole bude popísaný vybraný model pre predikciu sieťovej prevádzky.

6 Zmiešaný model ARIMA/GARCH

Na základe predchádzajúcej štúdií konkrétnych modelov spolu so zvážením referencií v literatúre [12] som zvolil zmiešaný ARIMA/GARCH model ako najvhodnejší k otestovaniu možností jeho využitia pri analýze a predikcii zachytenej sieťovej prevádzky. V tejto kapitole bude bližšie popísaný samotný model ARIMA/GARCH, a postup pri konštrukcii modelu. Nasledujúci text je súhrnom všetkých ‚doposiaľ prebratých, faktov, ktoré sa týkajú modelovania časových radov

6.1 Úvod k zloženému ARIMA/GARCH modelu

Už zo samotného názvu vyplýva, že sa jedná o kombináciu dvoch modelov lineárneho ARIMA(p,d,q) a nelineárneho modelu GARCH(p, q). Model môžeme označiť aj ako model podmienenej strednej hodnoty a podmieneného rozptylu. V nasledujúcich riadkoch si tento model vyjadríme matematicky. Aby nedošlo k zámene parametrov označujúcich rády jednotlivých modelov, u modelu ARIMA budeme používať namiesto parametra p parameter r, označujúci rád autoregresného procesu (AR). Označenie rádu q procesu kĺzavých súčtov (MA) nahradíme premennou m. Označenie rádu diferenciácie d a rádov procesu GARCH p a q necháme bez zmeny. Zápis ARIMA modelu bude nasledovný ARIMA(r, d, m).

Nech y_t sú série, ktoré chceme modelovať, potom všeobecný model **podmienenej strednej hodnoty** (ARIMA(r, d, m)) vyjadruje nasledujúci zápis:

$$(\nabla^d y_t) = \sum_{i=1}^{r} \phi_i (\nabla^d y_{t-i}) + a_t + \sum_{j=1}^{m} \theta_j (\nabla^d y_{t-j}), \tag{6.1}$$

Pripomeňme si, že miesto zápisu (6.1) môžeme použiť aj tvar (viď. podkapitola 2.3.4):

$$\phi_p(\mathrm{B})\Delta^d y_t = \theta_q(\mathrm{B})\, a_t$$

kde $a_t \sim N(0,\sigma_t^2)$, $\phi(.)$ a $\theta(.)$ predstavujú polynómy r-tého a m-tého stupňa, B je operátor spätného posunutia ($\mathrm{B}^j y_t = y_{t-j}$, $\mathrm{B}^j a_t = a_{t-j}$, $j = 0, \pm 1, \pm 2, \ldots$), hodnota d hovorí o stupni diferenciácie. Premenná a_t označuje inovácie originálneho časového radu.

$$\nabla y_t = y_t - y_{t-1} = (1-B) y_t \tag{6.2}$$

Vzťah (6.2) definuje diferenčný operátor s oneskorením „lag“ = 1.

Teraz definujme **podmienený rozptyl** inovácií a_t, pre ktorý platí [12]:

$$\sigma_t^2 = v_{t-1}(y_t) = E_{t-1}(a_t^2) \tag{6.3}$$

V pojednávaní o ARCH(p) procese (podkapitola 2.4.1) sme uviedli, že podstata jeho použitia spočíva v schopnosti rozlíšiť podmienený a nepodmienený rozptyl v procese inovácií a_t. Termín „podmienený“ jednoznačne naznačuje závislosť na predchádzajúcich pozorovaniach. Zatiaľ čo termín „nepodmienený“ sa týka dlhodobého chovania sa časového radu. Takéto chovanie si nevyžaduje zvláštne poznatky o minulých hodnotách radu. GARCH model teda na charakterizuje podmienenú distribúciu inovácií zachytením sériovej závislosti.

Všeobecný GARCH(p, q) model pre podmienený rozptyl inovácií a_t je,

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \alpha_i a_{t-i}^2 , \tag{6.4}$$

S obmedzeniami:

$$\sum_{i=1}^{p}\beta_i+\sum_{j=1}^{q}\alpha_i<1,\ \ \alpha_0>0\,,\ \alpha_i\geq 0\,,\ \beta_i\geq 0\,,\ i=1,2,...,q\,,\ i=1,2,...,p\,- \qquad (6.5)$$

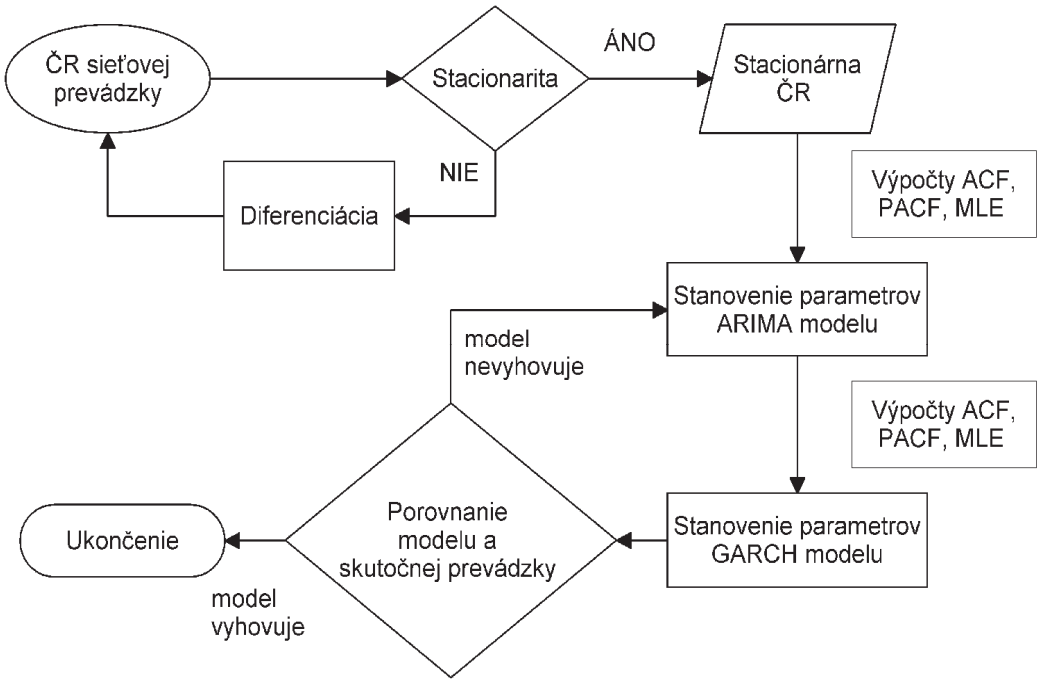
Kombináciou vzťahov (6.1) a (6.4) dostaneme matematické vyjadrenie celého ARIMA(r,d,m)/GARCH(p, q) modelu:

$$\begin{aligned}(\nabla^d y_t)&=\sum_{i=1}^{r}\phi_i(\nabla^d y_{t-i})+a_t+\sum_{j=1}^{m}\theta_j(\nabla^d y_{t-j}),a_t\sim N(0,\sigma_t^2)\\ \sigma_t^2&=\alpha_0+\sum_{i=1}^{p}\beta_i\sigma_{t-i}^2+\sum_{j=1}^{q}\alpha_i a_{t-i}^2\end{aligned} \qquad (6.6)$$

6.2 Odhad parametrov

Odhad parametrov modelu je prvým krokom jeho prispôsobenia analyzovaným dátam. Celý postup odhadu, celkovo stanovenie piatich parametrov r, d, m, p, q vychádza z Box-Jenkinsovej metodológie konštrukcie modelu (kapitola 3).

Postup odhadu parametrov stručne zachytáva vývojový diagram (Obrázku č.2).



Obr. č. 2: *Vývojový diagram popisujúci odhad parametrov modelu* ARIMA(r,d,m)/GARCH(p, q).

Ako prvý krok volíme stanovenie parametru d vyžaduje si to podmienka stacionarity, ktorá predurčuje konštrukciu uspokojivého modelu. Ak sa časový rad analyzovanej prevádzky správa nestacionárne musíme ju diferencovať. Parameter d odhadneme na základe výpočtu ACF a PACF (viď. podkapitola 3.1).

ARIMA model popisuje podmienenú strednú hodnotu, ktorá sa mení v čase avšak pri konštantnom rozptyle pozorovaných dát. Rád parametrov r a m určíme taktiež pomocou vlastností ACF a PACF. Parameter m procesu MA môžeme priamo určiť

z ACF, ktorá popisuje koreláciu medzi aktuálnymi a minulými hodnotami časovej rady. PACF, ktorá charakterizuje koreláciu medzi aktuálnymi hodnotami a ich zmenou oproti minulosti, je vhodná na určenie parametra *r* autoregresného procesu.

Parametre p a q ,nelineárneho modelu, sa klasickým spôsobom určujú pomocou autokorelačnej a parciálnej autokorelačnej funkcie podobne ako u ARIMA modelu. Na základe ACF a PACF však tieto parametre nie sme schopný určiť presne. Z praxe je známe , že nastavenie týchto dvoch parametrov na hodnotu 1 je adekvátnym riešením z hľadiska zachytenia podmieneného rozptylu. Znamená to, že by sme si mali vystačiť s nelineárnym modelom GARCH(1, 1).

Po odhadnutí parametrov *r*, *d*, *m*, *p*, *q* modelu nasleduje výpočet reálnych hodnôt parametrov $\phi_r, \theta_m, \alpha_q, \beta_p$. Aby nedošlo k omylu budeme tieto parametre označovať pojmom koeficienty. Samotný proces výpočtu týchto koeficientov je oblasťou numerickej matematiky. Používajú sa rôzne iteračné metódy. Často využívanou metódou je metóda maximálnej pravdepodobnosti (MLE) [18]. Postup stanovenia koeficientov ϕ_r, θ_m je nasledovný [12]: Najskôr sa vypočítajú AR koeficienty ϕ_r pomocou výpočtu autokovariančnej matice a vyriešením Yule-Walkerových rovníc. Následne sú tieto koeficienty využité pri filtrácii pôvodného časového radu, čím sa získa čistý MA proces. Koeficienty θ_m MA procesu sa potom iteratívnou metódou stanovia z vopred vypočítanej autokovariančnej sekvencie tohto procesu.

Čo sa modelu GARCH týka, budeme sa držať podmienky z (6.5): $\sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_i < 1$,

pre koeficienty $\sum_{j=1}^{q} \alpha_i$ môžeme alokovať hodnotu približne 0,05 a pre koeficienty $\sum_{i=1}^{p} \beta_i$ alokujeme hodnotu 0,85. V praxi by takého stanovenie malo byť dostačujúce. Keďže bolo zmienené, že si v praxi vystačíme s modelom GARCH(1,1), môžeme model podmieneného rozptylu rovno popísať vzťahom:

$$\sigma_t^2 = k + 0,85\sigma_{t-1}^2 + 0,05\alpha_{t-1}^2 \tag{6.7}$$

kde konštantu *k* získame zo vzťahu pre výpočet nepodmieneného rozptylu zložky a_t:

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T} a_t^2 \quad . \tag{6.8}$$

Vzťah (6.8) môžeme vyjadriť:

$$\sigma^2 = \frac{k}{1 - \sum_{i=1}^{p} \beta_i - \sum_{j=1}^{q} \alpha_j} = \frac{k}{1-(0,85+0,05)} \text{ [12]} \tag{6.9}$$

z toho vyplýva, že :

$$k = \sigma^2(1-(0,85+0,05)) = 0,1\sigma^2 . \tag{6.10}$$

7 Praktická časť

Táto kapitola dokumentuje vlastnú prácu, ktorá vychádza z pojednávanej teórie v predchádzajúcich kapitolách. Bude v nej popísaná fáza prípravy a spracovania dát sieťovej prevádzky a napokon aj samotná analýza a postup stanovenia modelu vybranej prevádzky.

7.1 Príprava a spracovanie série pozorovaných dát

Pre samotnú analýzu budú použité dáta získané zo záznamu relatívne vyťaženého HTTP servera s doménou www.ukf.sk. Zdrojom dát je záznam z 10. až 19. Januára 2012 uložený vo formáte „.log“, ktorý som získal od študenta UKF v Nitre. Formát „.log“ je všeobecný štandardizovaný textový formát obsahujúci záznamy podľa pevnej syntaxi „*host ident authuser date request status bytes*“

Zo záznamov z uvedeného obdobia som získal vybraný 48 hodinový záznam z dní 17. (utorok) a 18. (streda) Januára 2012. Výber bol realizovaný pomocou freeware utility *Wingrep*.

Pre analýzu prevádzky boli vybrané iba niektoré kľúčové informácie a to: **časy príchodov požiadaviek**, **veľkosť vyžiadaného súboru v bytoch**, **prístupy na hlavnú stránku** (prístup na domovskú stránku z dôvodu uvažovania, že ide vytvorenie nového spojenia, respektíve sedenia).

Z dôvodu optimalizácie rýchlosti spracovania som vybraný záznam očistil od prebytočných údajov. Táto selekcia bola spracovaná pomocou freeware utility *CSV-editor* (verzia 2.2.0). Takto očistené, chronologicky zoradené, dáta sú priložené v prílohe A, v komprimovanej zložke s názvom „17-18_jan_2012_48h_log_on_server_ukf.sk.zip”, jednotlivé dáta sú uložené v textových súboroch.

Ďalším krokom vo fáze prípravy dát bolo vyriešenie problému okamihovej časovej rady, ktorá musela byť transformovaná na intervalovú časovú radu. Tento problém vystihuje Tabuľka 1, kde sa v ľavom stĺpci nachádza úsek dát s časmi príchodu požiadaviek vo formáte *HH:MM:SS* a na pravej strane je výsledná stanovená časová rada s intervalom pozorovania 30 sekúnd. V princípe ide o jednoduchý počítací proces.

časový údaj HH:MM:SS
00:18:03
00:18:04
00:18:05
00:19:13
00:19:40
00:20:08

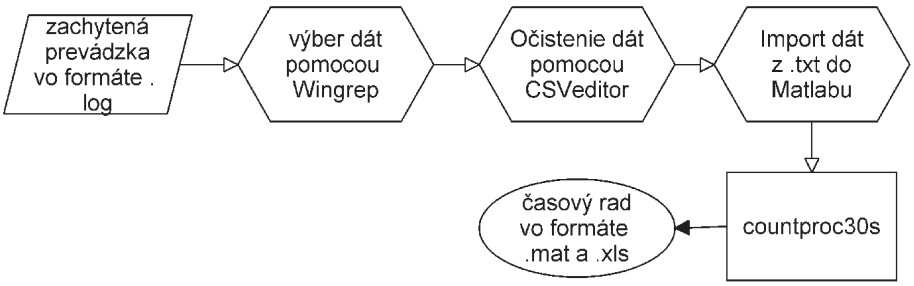
countproc30s →

časový rad **yt**	
t (30 s)	počet požiadaviek
37	3
38	0
39	1
40	1
41	1

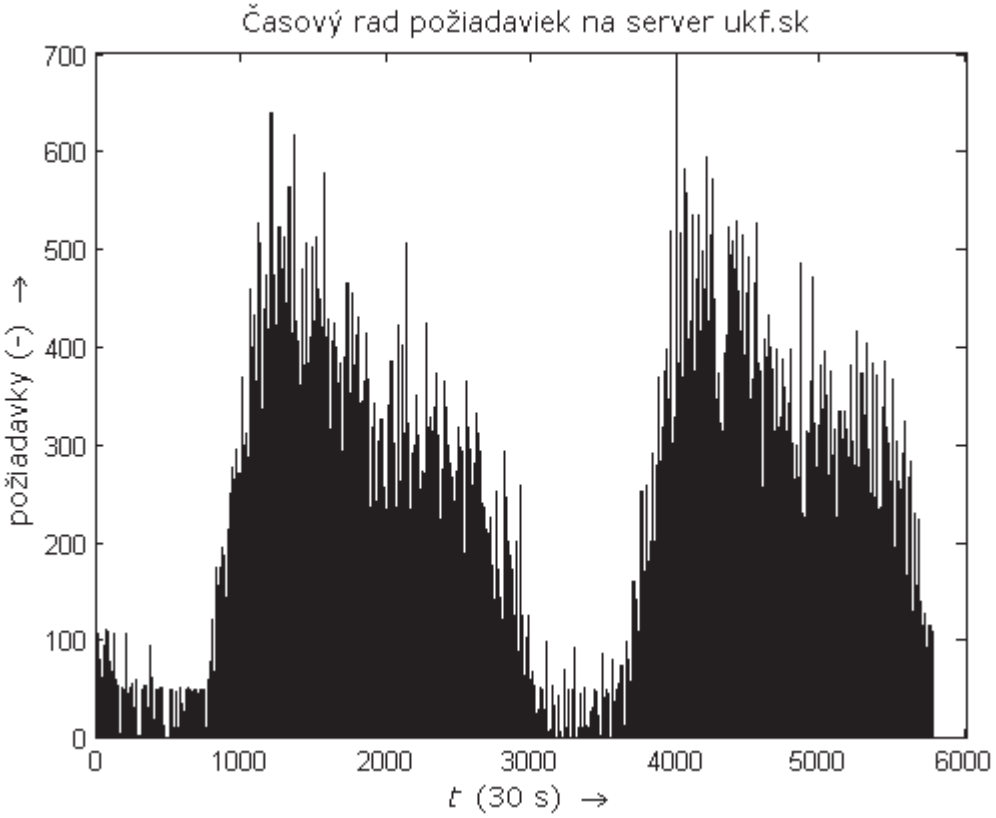
Tabuľka č. 1: *Transformácia časových údajov na časový rad (časové údaje môžeme chápať ako okamihový časový rad)*

Demonštrovaná transformácia (Tabuľka 1) bola prevedená pomocou funkcie *countproc30s* napísanej v jazyku Matlab. Kód je priložený v prílohe B . V prílohe A sú priložené aj upravené dátové štruktúry so zachytenou prevádzkou, ktoré sú vstupnými dátami funkcie. Výstupom funkcie *countproc30s* sú časové rady s 30 sekundovým

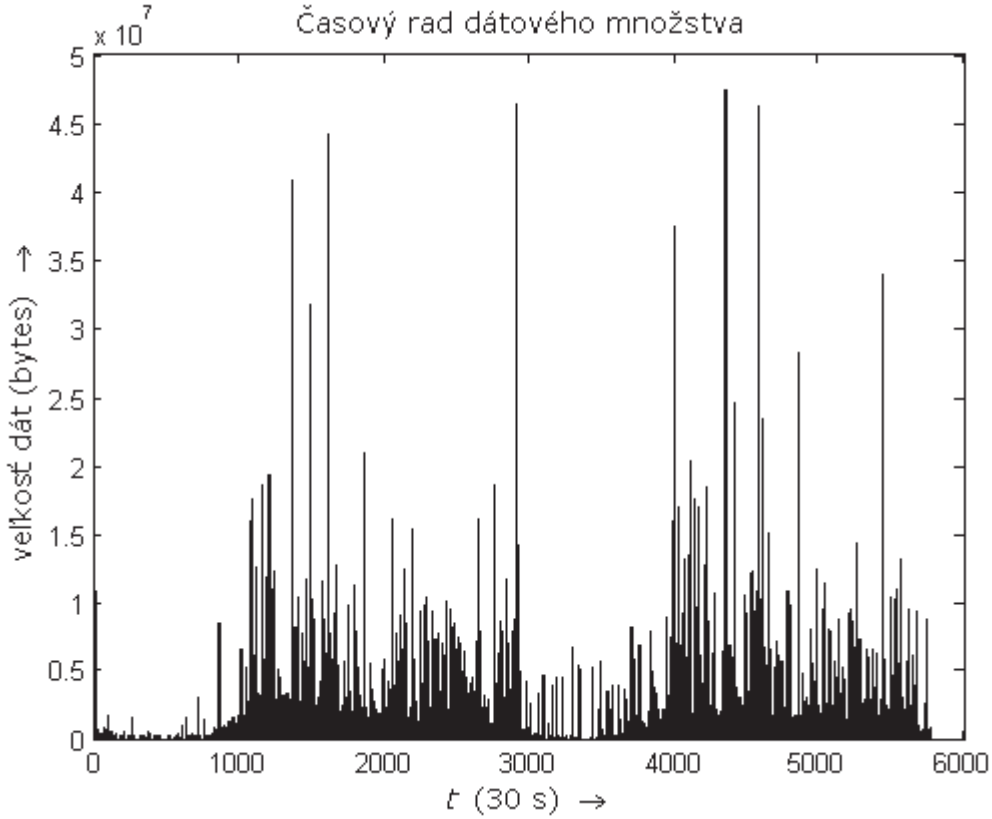
intervalom pozorovania. Ich priebeh je graficky znázornený na Obr. č. 4, Obr. č. 5 s Obr. č. 6. Hodnoty časových radov sú taktiež priložené v prílohe A.



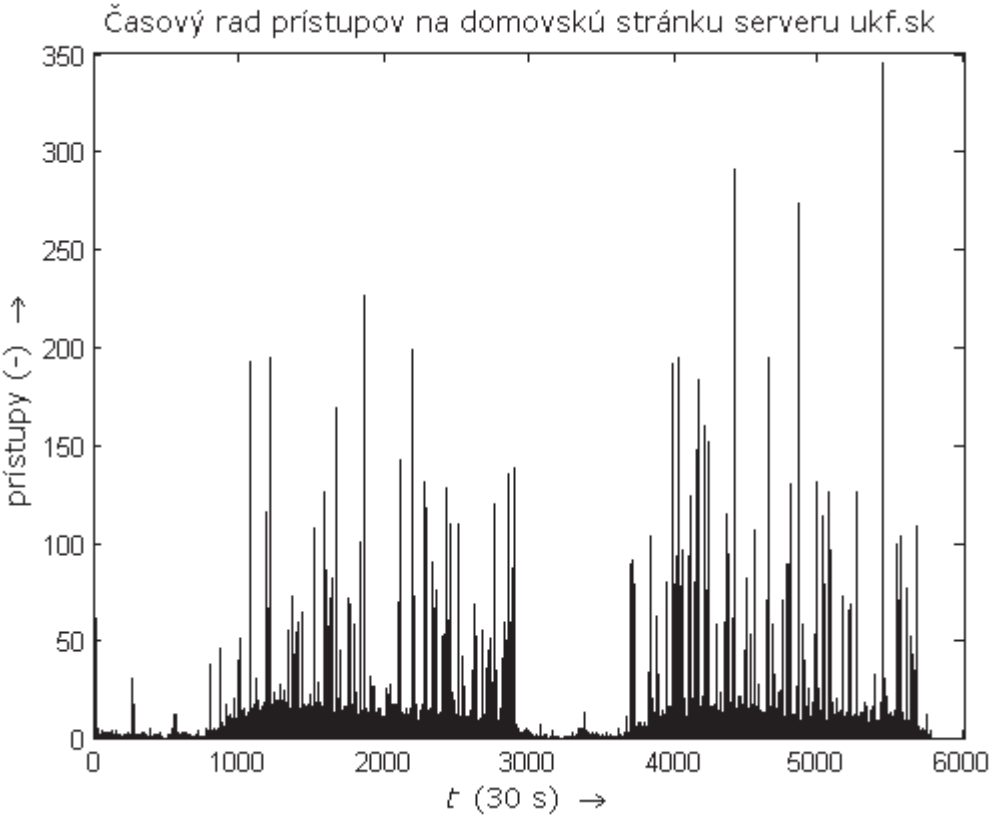
Obr. č. 3: *Vývojový diagram popisujúci prípravu a predspracovanie dát.*



Obr. č. 4: *Časový rad požiadaviek na server*



Obr. č. 5: *Časový rad množstva vyžiadaných dát*

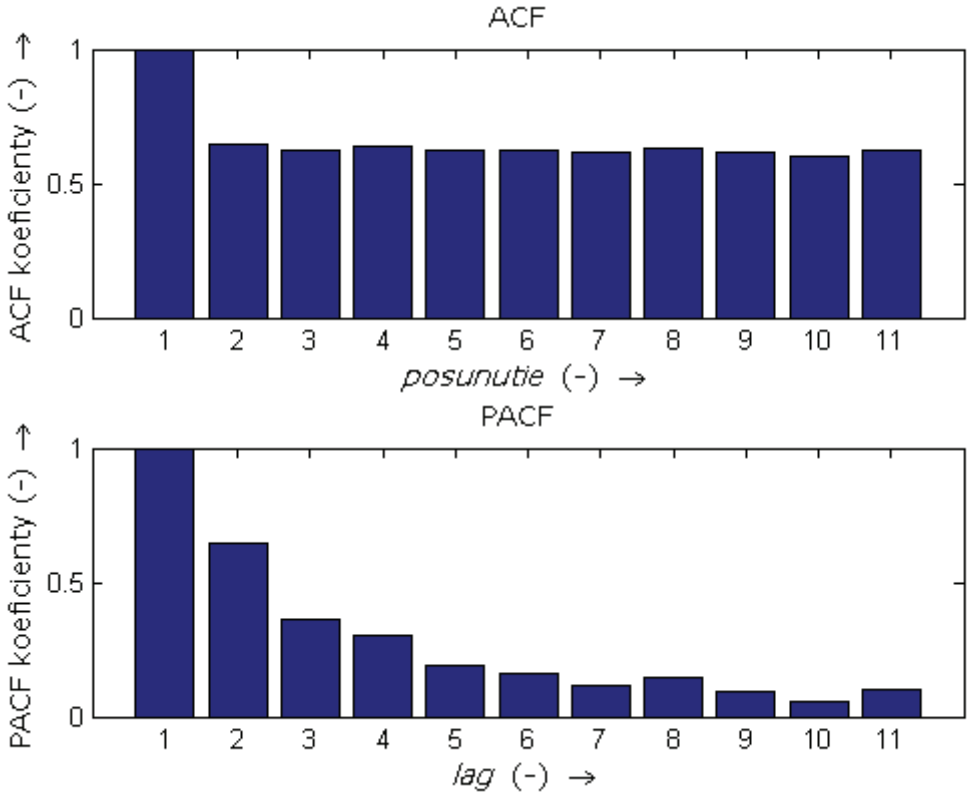


Obr. č. 6: *Časový rad množstva vyžiadaných dát*

7.2 Konštrukcia modelu prevádzky

Konštrukcia modelu vychádza z postupov uvedených v kapitole 6. Analýza a prispôsobenie modelu bude aplikovaná na časový rad požiadaviek (Obr. č. 4).Postup čiastočne dokumentuje skript s názvom *arima_garch.m* (viď. príloha B). Pri pokuse o stanovenie modelu boli využité špeciálne funkcie z ekonometrickej sady nástrojov prostredia Matlab.

Po načítaní časových radov pomocou funkcií *atocorr()* a *parcorr()* dostávame ACF a PACF. Ich priebeh naznačuje, že ide o nestacionárny časový rad a je potrebné ho stacionarizovať diferencovaním.

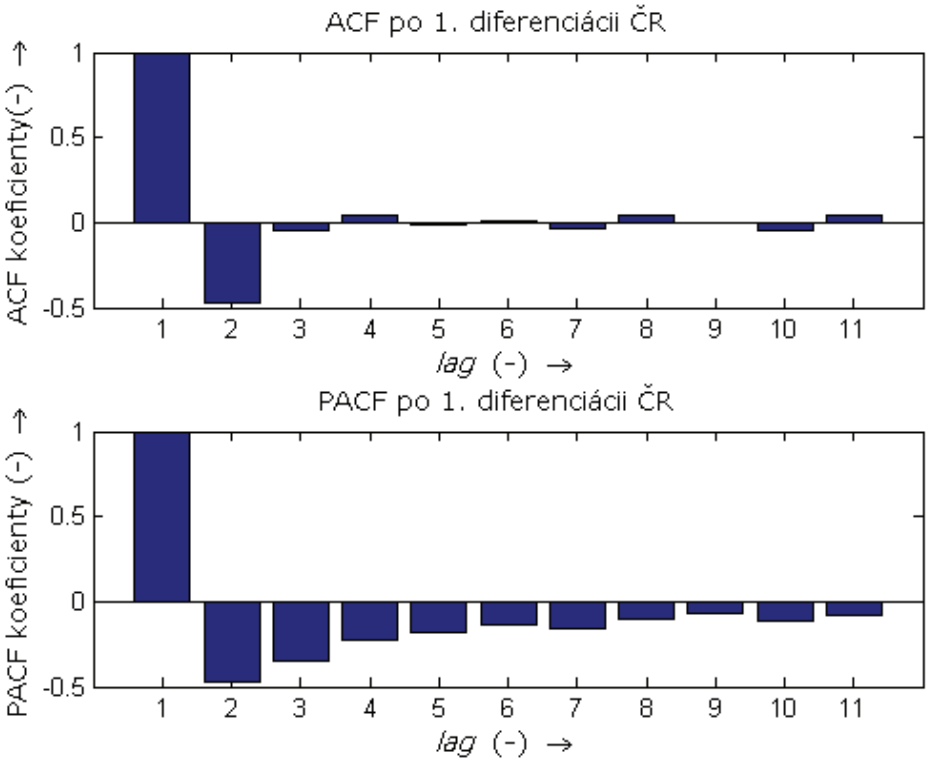


Obr. č. 7: *Priebeh ACF a PACF časového radu požiadaviek na server ukf.sk*

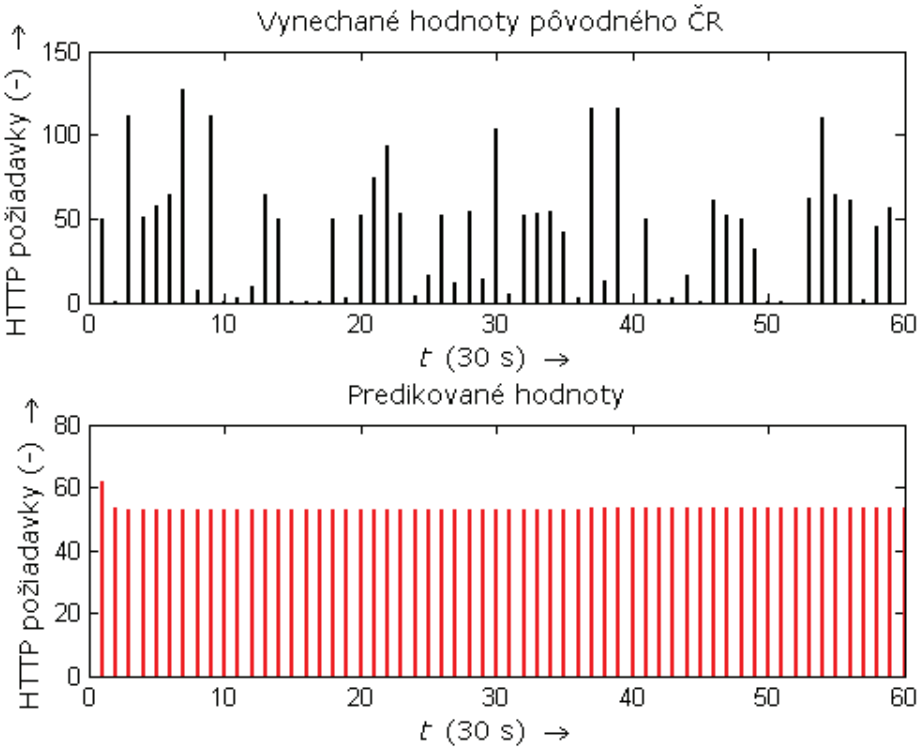
Skript využíva funkciu *garchspec()* bližšie popísanú v prílohe B a v jej dokumentácií, ktorá vytvorí štruktúru *Spec* obsahujúcu explicitne zadané hodnoty parametrov zmiešaného modelu. Táto štruktúra môže byť následne spolu s časovým radom využitá ako vstup ďalšej funkcie *garchfit()*, ktorá na základe explicitnej špecifikácie v *Spec* modelu vypočíta koeficienty danému časovému radu, t. j. vypočíta hodnoty koeficientov a uloží do novej štruktúry *Coeff*.

Napriek dlhšiemu skúmaniu dokumentácie, som neprišiel na to ako explicitne vložiť do tejto štruktúry hodnotu parametra d , t.j. zaviesť diferenciáciu. Hoci hodnoty parametrov *p, q, r, m,* a taktiež koeficientov čiastkových modelov AR, MA, ARCH, GARCH je veľmi jednoduché explicitne zadať viď. dokumentácia funkcie. Rozhodol som sa teda časový rad diferencovať pomocou funkcie *diff()*. Priebehy ACF a PACF po diferenciácií sú graficky zobrazené na Obr. č. 8. Podľa priebehov môžeme usúdiť, že časový rad je stacionarizovaný. Po overení stacionarity nasleduje explicitné vytvorenie štruktúry *Spec* pomocou zmienenej funkcie a následne volanie funkcie *garchfit()*.

Usudzujem, že tým by mal byť proces odhadu koeficientov ukončený. Vytvorením predpovedí na základe modelu a pozorovanej časového radu (funkcia *garchpred()*) som dospel k záveru, že stanovený model je nevhodný vid Obr. č.9.



Obr. č. 8: *Priebeh ACF a PACF po diferenciácii*



Obr. č. 9: *Priebeh ACF a PACF po diferenciácii*

Na prvý pohľad model nepredikuje variabilitu. Preto som sa rozhodol vyskúšať viacero variant explicitným stanovením modelu, avšak bezúspešne. Táto skutočnosť ma prinútila k rôznym úvahám. Napríklad je možné, že v 48 hodinovej prevádzke sa vyskytuje systematická zložka v podobe sezónnosti, ktorú nie je schopný daný model popísať.

Záver

Práca pojednáva o problematike modelovania a predikcie prevádzky v dátových sieťach a možnostiach analýzy pomocou časových radov. Prvé stránky sa venujú teórii časových radov, ich vlastnostiam a postupom pri konštrukcií modelov. Podrobnejšie sú popísané lineárne modely AR, MA, ARMA, ARIMA a taktiež nelineárne modely ARCH a GARCH.

V kapitole 4 sú spomenuté niektoré konvenčné modely, a ich neschopnosť zachytiť niektoré špecifické vlastnosti súčasnej sieťovej prevádzky. Kapitola 5 rozoberá a vysvetľuje charakter HTTP prevádzky. Definuje pojmy ako sú sebe-podobosť, či dlhodobá závislosť.

Na základe štúdia modelov časových radov a referencií z použitej literatúry bol vybraný konkrétny model ARIMA/GARCH, pomocou ktorého som sa pokúsil modelovať a následne predikovať zachytenú sieťovú prevádzku. ARIMA/GARCH model všeobecne je bližšie špecifikovaný v kapitole 6. V kapitole 7. Praktická časť je zdokumentovaná vlastná práca. Opisuje použité dáta zachytenej sieťovej prevádzky, najmä ich prípravu, ktorá mala za cieľ vydolovať z nich pre nás zaujímavé informácie a pretransformovať ich na časové rady s intervalom. Z dát boli takto vytvorené celkovo 3 časové rady s intervalom pozorovania 30 s. Po príprave dát nasledovala konštrukcia modelu: odhadnutie, stanovenie parametrov a predikcia. Táto časť bola realizovaná v prostredí Matlab pomocou jeho štatistických nástrojov. Na základe vizuálneho posúdenia z dokumentujúcich grafov (kapitola 7), sa mnou stanovený model neosvedčil. Predpokladám však, že hlbšie štúdium čiastkových procesov, vedúcich k stanoveniu vhodného modelu, spolu s rozšírením o ďalšie testovacie metódy by mohlo dospieť k priaznivým výsledkom. Samozrejme nemôžem vylúčiť ani existenciu vhodnejšieho prostredia, ktorý by bol adekvátny na realizáciu cieľov tejto práce.

Napriek môjmu neúspechu, vďaka štúdiu teórie som stále presvedčený, že modely časových radov majú veľký potenciál aj v oblasti modelovania sieťovej prevádzky, a sú schopné uspieť tam kde iné konvenčné modely zlyhávajú. Nesporným dôkazom o ich potenciáli, je aj fakt, že množstvo týchto modelov je využívaných v rôznych priemyselných oblastiach, nehovoriac o širokom uplatnení v ekonometrii.

Literatúra

[1] WILSON, Michael *A historical view of network traffic models* [online]. 2006. Dostupný také z WWW: <http://www.cs.wustl.edu/~jain/cse567-06/ftp/traffic_models2.pdf>.

[2] BECCHI, Michela. *From Poisson Processes to Self-Similarity: a Survey of Network Traffic Models* [online]. 2006. Dostupný také z WWW: <http://www1.cse.wustl.edu/~jain/cse567-06/ftp/traffic_models1/index.html>.

[3] MOLNÁR, Karol. *Moderní síťové technologie* [online]. Brno : FEKT VUT Brno, 2008. Dostupný také z WWW: <http://www.utko.feec.vutbr.cz/~molnar/mmos/fronty.pdf>.

[4] CROVELLA, Mark E.; BESTAVROS, Azer. Self-Similarity in World Wide Web Traffic : Evidence and Possible Causes [online]. *In IEEE/ACM Transactions on Networking*. December 1997, Vol 5, Number 6, s. 835-846. Dostupný také z WWW: <http://users.cms.caltech.edu/~adamw/courses/147/2009/lectures/ton97.pdf>.

[5] WILLINGER, Walter, et al. Self-Similarity in High-Speed Packet Traffic : Analysis and Modeling of Ethernet Traffic Measurements [online]. *Statistic Science*. February, 1995, Vol 10, Issue 1, s. 67-85. Dostupný také z WWW: <http://www.cs.northwestern.edu/~srg/Papers/Queue/willinger-self-similar.pdf>.

[6] ARLT, Josef; ARLTOVÁ, Markéta; RUBLÍKOVÁ, Eva. *Analýza ekonomických časových řad s příklady : Skripta VŠE Praha*. 2. vyd. Praha : [s.n.], 2004. 148 s. Dostupné z WWW: <http://nb.vse.cz/~arltova/vyuka/crsbir02.pdf>. ISBN 80-245-0777-3.

[7] KŘIVÝ, Ivan. *Analýza časových řad* [online]. Ostrava, 2006. 77 s. Dostupné z WWW:<http://www.informatika-osu.czechian.net/files/is/ancas/ANCAS_DiV.pdf>

[8] FAJMON, Břetislav, RŮŽIČKOVÁ, Irena. *MATEMATIKA_3_S.PDF* [online]. Matematika 3. Brno: UMAT FEKT VUT, 2003. s. 1-266. Dostupné z WWW: <http://www.umat.feec.vutbr.cz/~hlavicka/skripta/matematika3.pdf>

[9] ANDĚL, Jan. *Statistická analýza časových řad*. Praha : SNTL, 1976.

[10] CIPRA, Tomáš.: *Analýza časových řad s aplikacemi v ekonomii*. SNTL. Praha. 1986. s. 246.

[11] PERELLI, Roberto. *Introduction to ARCH & GARCH models* [online]. [s.l.], 2001. 7 s. Optional TA Handout. University of Illinois. Dostupné z WWW: <http://www.econ.uiuc.edu/~econ472/ARCH.pdf>.

[12] ZHOU, Bo, et al. *Network Traffic Modeling and Prediction with ARIMA/GARCH* [online]. [s.l.] : [s.n.], [200?]. s. 10. Dostupné z WWW: <http://www.econ.uiuc.edu/~econ472/ARCH.pdf>.

[13] LELAND, W.E., et al. On the Self-Similar Nature of Ethernet Traffic[online]. *SIGCOMM 93*. 1993, Vol. 23, No. 4, s. 1-12. Dostupný také z WWW: <http://netlab.caltech.edu/FAST/references/ccr-9501-leland.pdf>.

[14] BORCHERS, Brian. *Notes on ARIMA Modelling* [online]. [s.l.], 2002. 19 s. Poznámky. X. Dostupné z WWW: <http://panda.bg.univ.gda.pl/~prezes/pliki/matlab/arima.pdf>.

[15] HUNG,Bill. Dickey-Fuller Unit Root test [online]. [200?] [cit. 2011-12-15]. Dickey-Fuller Unit Root Test. Dostupné z WWW: <http://www.hkbu.edu.hk/~billhung/econ3600/application/app01/app01.html>.

[16] Bollerslev, T. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*. Vol. 31, 1986, pp. 307–327.

[17] SOUKAL, Petr. *Heteroskedasticita*. Praha: Vysoká škola ekonomická Praha, Fakulta informatiky a statistiky,1999. 103 s. Vedoucí práce Prof. Ing. Petr HEBÁK, CSc
[18] JAE MYUNG, In. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*.[online] 2003, č. 47. Dostupné z WWW: <http://people.physics.anu.edu.au/~tas110/Teaching/Lectures/L3/Material/Myung03.pdf>
[19] Cvičenia z časových radov. BEÁTA STEHLÍKOVÁ. [online]. Dostupné z WWW: <http://www.iam.fmph.uniba.sk/institute/stehlikova/cr09.html>

Zoznam použitých skratiek

ACF – Autocorrelation Function (autokorelačná funkcia)

AR – Autoregressive model (autoregresný model/proces)

ARMA – Autoregressive-Moving Average model (zmiešaný model/proces AR a MA)

ARIMA – Autoregressive Integrated Moving Average model/proces
(zmiešaný model AR integruje MA)

ARCH – Autoregressive Conditional Heteroskedasticity model/process
(autoregresívny podmienený heteroskedastický model/proces)

ČR – Časový Rad

EPA – Environmental Protection Agency (agentúra ochrany životného prostredia)

GARCH – Generalized Autoregressive Conditional Heteroskedasticity model/process
(generalizovaný autoregresný podmienený heteroskedastický model/proces)

HTTP – Hyper-text Transfer Protocol
(hypertextový transportný protokol)

IID – Independent Identically Distributed
(nezávislé identické rozdelenie)

ITA – Internet Traffic Archive (archív internetovej prevádzky)

LRD – Long Range Dependency (dlhodobá závislosť)

MA – Moving Average (model/proces kĺzavých súčtov)

NMA – Nonlinear Moving Average
(nelineárny proces kĺzavých priemerov)

PACF – Partial Autocorrelation Function (parciálna autokorelačná funkcia)

SRD – Short Range Dependency (krátkodobá závislosť)

Zoznam príloh

Príloha A: CD s dátami:

countproc30s.m	funkcia spracúvajúca dáta v jazyku Matlab
arima_garch.m	skript v jazyku Matlab realizujúci konštrukciu modelu
ts.xls	časové rady požiadaviek (1.hárok), bytov (2.hárok), prístupov na domovskú stránku (3.hárok) vytvorené s intervalom pozorovania 30min
17-18_jan_2012_ 48h_log_on_server _ukf.sk.zip	vybrané a upravené záznamy servera ukf.sk z obdobia 17. až 18. Január 2012, uložené v textových súboroch .txt
time.mat, bytes.mat, home.mat	štruktúry obsahujúce z .txt do matlabu importované očistené a upravené informácie zachytenej sieťovej prevádzky, dáta slúžia ako vstupy funkcie *countproc30s.m*
yt.mat, yb.mat yh.mat	výstupné dáta funkcie *countproc30s.m*, t.j. sieťová prevádzka prevedená na časové rady pripravené na analýzu.

Príloha B: Zdrojový kód funkcie *countproc30.m* a kód skriptu *arima_garch.m*.

Príloha B

Zdrojový kód funkcie countproc30s

```
%%%%%%%%%%%%%%%% JAN PAUKEJE, VUT BRNO. 2011%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% POMOCNA FUNKCIA, KTORA MA ZA ULOHU TRANSFORMOVAT OKAMIHOVY CASOVY RAD
% NA INTERVALOVY S INTERVALOM POZOROVANIA 30 s.

function [yt yb yh] = countproc30s

clear all;
clc
close all;

% NACITANIE DAT

% nacitanie matice Mx3 casových udajov vo formáte
%(M,1)= hodina, (M,2)= minuta, (M,3)=sekunda
structxt = load('time.mat');
% nacitanie stlpcoveho vektora Mx1 s hodnotami velkosti dat v bytoch
structxb = load('bytes.mat');
% nacitanie stlpcového vektora Mx1, kde hodnota 1 identifikuje pristup
na
% domovsku stranku, v opacnom pripade 0.
structxh = load('home.mat');


% EXTRAHOVANIE MATICE S CASOVYMI UDAJMI PRICHODU POZIADAVIEK ZO
STRUKTURY
xt = getfield(structxt,'time');
xb = getfield(structxb,'bytes');
xh = getfield(structxh,'home');


% INICIALIZACIA PREMENNYCH c - POCITADLO POZIADAVIEK po 1 s, c30 -
POCITADLO
% POZIADAVIEK po 30 S , yt - TS VEKTOR RADU,i - POMOCNA PREMENNA PRI
% ITERACIACH, t - pocitadlo sekund, t30 - POCITADLO po 30 s

ct=0;
ct30=0;
cb=0;
cb30=0;
ch=0;
ch30=0;


t=0;
t30=0;

yt=zeros(5760,1); % 48h x 60m x 2(30s) = 5760 polminutovych udajov
yb=zeros(5760,1);
yh=zeros(5760,1);
i=1;
n=length(xt)
```

```
% PREDLZI VEKTOR X O 1, KVOLI SPRAVNEMU UKONCENIU SLUCKY WHILE
xt(n+1,1) = 0;
xb(n+1,1) = 0;
xh(n+1,1) = 0;

% INICIALIZACIA POMOCNYCH PREMENNYCH OPERUJUCICH S CASOVYMI UDAJMI
sS = xt(1,3);
aS = xt(1,3);
sM = xt(1,2);
aM = xt(1,2);
sH = xt(1,1);
aH = xt(1,1);

t = sS;

% SLUCKA 1 WHILE BUDE BEZAT NAD CELOU DLZKOU VEKTORA CASOVYCH UDAJOV
while (i <= n)

    while (t <= 29)
%     PODMIENKA 'HODIN' MA ZARUCIT STABILITU FUNKCIE V PRIPADE, SKAKANIA
%     CASOVYCH UDAJOV, NAPRIKLAD AK ZA CASOVYM UDAJOM 1:11:1 NASLEDUJE
UDAJ
%     1:12:1, KDE SA UDAJ O SEKUNDE ROVNA. FUNKCIA SA TAK ZISTI, ZE SA
JEDNA
%     O SEKUNDU NASLEDUJUCEJ MINUTY A NEBUDE POKRACOVAT V POCITANI.
        while (sS == aS)&&(sM == aM)&&(sH == aH)

    ct=ct+1;

    cb=cb+xb(i,1);
    ch=ch+xh(i,1);
    i=i+1;

    aS=xt(i,3);
    aM=xt(i,2);
    aH=xt(i,1);

        end

% "DIGITALNE HODINY FUNKCIE", STARAJU SA O SPRAVNY ITERACNY CYKLUS
% HODIN, MINUT A SEKUND. RESPEKTIVE OBMEDZUJU MAXIMALNU HODNOTU MINUT
% A SEKUND NA 59, PO HODNOTE ...59 NASLEDUJE OPAKOVANIE 0,1,...
    if sS<59
        sS=sS+1;
    else sS=0;
        if sM<59
            sM=sM+1;
         else sM=0;
             if sH<23
                 sH=sH+1;
             else sH=0;
             end
        end
    end

    t=t+1; % pocitadlo +1s
    ct30=ct30+ct; %pocitadlo poziadaviek po 30 sekundach, + pocet za 1s
    ct=0; % nulovanie pocitadla poziadaviek 1s
```

```
    cb30=cb30+cb;
    cb=0;
    ch30=ch30+ch;
    ch=0;

    end

    t=0; % nulovanie pocitadla t, zacne sa pocitanie novych 30
    t30=t30+1; %pocitadlo casu po 30 s, +1x30 interval
    yt(t30,1)=ct30; % vlozenie poctu za interval 30 s do vektora yt
    yb(t30,1)=cb30;
    yh(t30,1)=ch30;
    ct30=0; %nulovanie
    cb30=0;
    ch30=0;
end

% ULOZENIE VYSTUPNYCH TS DO MATLAB SUBOROV
save yt yt
save yb yb
save yh yh

% MOZNOST EXPORTU UDAJOV PRIAMO DO EXCEL SUBORU .XLS
% POZOR! EXCEL NEPRIJME VACSI POCET RIADKOV AKO 2^16=65536

xlswrite('ts.xls',yt(:,1),1); % 1. harok pocty poziadaviek
xlswrite('ts.xls',yb(:,1),2); % 2. harok pocet bytov
xlswrite('ts.xls',yh(:,1),3); % 3. harok pocet home page
```

Zdrojový kód skriptu arima_garch

```
%%%%%%%%%%%%%%%%%%%%%%JAN PAUKEJE, VUT BRNO, 2012.%%%%%%%%%%%%%%%%%%%%

% Skript realizujúci konštrukciu ARIMA/GARCH modelu. Využíva funkcie
% obsiahnuté v Ekonometrickej a Štatistickej sade nástrojov prostredia
% MATLAB (R2009b).


%% Čistka

%Zmazanie príkazového riadku.Zatvorenie všetkých okien. Odstránenie
%premenných z pracovného priestoru.
clc;
clear all;
close all;
%% Načítanie dát

% Načítanie stĺpcového vektora obsahujúceho časové rady v poradí:
% 1. prichádzajúce požiadavky na server yt.mat
% 2. veľkosť vyžiadaného súboru v bytoch yt.mat
% 3. prístupy k domovskej stránke yh.mat
% Všetky 3 časové rady sú vytvorené s intervalom pozorovania 30s.

structy = load('yt.mat'); % prichádzajúce požiadavky na server,
y = getfield(structy,'yt');

% structy = load('yb.mat'); % vyžiadaná veľkosť dát v bytoch
% y = getfield(structy,'yb')
%
% structy = load('yh.mat'); % prístupy k domovskej stránke.
% y = getfield(structy,'yh');

trainSeries = y(1:5701,1); % orezanie spočívajúce vo vynechaní
posledných
                           % 60 hodnôt kvôli testovaniu s predikovanými
                           % hodnotami

fullSeries = y;            % celý vektor časového radu

%% Výpočet ACF a PACF a diferenciácia, výpočet strednej hodnoty

lag = 10; % oneskorenie udáva počet koeficientov, ktoré majú byť
výstupom
          %korelačných funkcií
acf = autocorr(trainSeries, lag); % výpočet koeficientov ACF
pacf = parcorr(trainSeries, lag); % výpočet koeficientov PACF

trainSeries = diff(trainSeries);

dacf = autocorr(trainSeries, lag); % výpočet koeficientov ACF po
diferenciacii
dpacf = parcorr(trainSeries, lag); % výpočet koeficientov PACF


%% Špecifikácia modelu
```

```
% Táto funkcia vytvorí štruktúru, ktorá bude obsahovať odhad budúceho
% modelu, respektíve základnú špecifikáciu, ktorej by mala predchádzať
% identifikácia modelu. Jej výstup Spec je v nasledujúcom riadku kódu
% použitý ako vstup funkcie garchfit().
% Jednotlivé znaky, za ktorými nasleduje číselná špecifikujú rád
% čiastkových procesov AR, MA, GARCH,ARCH, prípadne priamo hodnoty
% koeficientov modelu. Viď. dokumentácia funkcie.

Spec = garchset('R',1,'M',1,'C',0,'K',0.0001,'P',1,'Q',1);

%% Stanovenie parametrov modelu ARMAX/GARCH

%Funkcia garchfit() - jej vstupom je načítaný časový rad a štruktúra
Spec
%popísaná vyššie. Výstupom funkcie,ktorý nás obzvlášť zaujíma je
štruktúra
%Coeff obsahujúca základné informácie o stanovenom modeli spolu s
%odhadnutými parametrami modelu. Štuktúra Errors, ako napovedá jej
názov,
%obsahuje štandardné chyby odhadnutých parametrov.
% Táto funkcia vykoná prispôsobenie modelu časovému radu pomocou výpočtu
% neznámych parametrov, prípadne prepočet parametrov stanovených v
% predchádzajúcom kroku v štruktúre Spec.

[Coeff,Errors,LLF,Innovations,Sigmas,Summary] =
garchfit(Spec,trainSeries);

%% Simulácia
% Nasledujúca funkcia pomocou stanoveného modelu (popísaného v štruktúre
% Coeff) simuluje nový kvazi-náhodný časový rad. Voliteľná vstupná
premenná
% NumSamples definuje počet hodnôt, ktoré majú byť simulované.

NumSamples = 5760;
[simInnovations, simSigmas, simSeries] = garchsim(Coeff, NumSamples);

%% Predikcia
% Funkcia garchpred() na základe vypočítaných parametrov modelu
uložených v
% štruktúre Coeff, predikuje nové nastávajúce hodnoty časového radu.
% Premenná NumPeriods špecifikuje počet predikovaných hodnôt.

NumPeriods = 60;
[SigmaForecast, MeanForecast,...
    SigmaTotal, MeanRMSE] = garchpred(Coeff, trainSeries, NumPeriods)

%% Vykreslenie dát

figure(1)

subplot(2,1,1), bar(acf);

% Create xlabel
xlabel('{\itposunutie } (-)  \rightarrow','FontName','MS Reference Sans
Serif');
% Create ylabel
ylabel('ACF koeficienty (-)  \rightarrow',...
    'FontName','MS Reference Sans Serif');
```

```
% Create title
title('ACF {\it} ',...
    'FontName','MS Reference Sans Serif');

subplot(2,1,2), bar(pacf);

% Create xlabel
xlabel('{\itlag } (-)  \rightarrow','FontName','MS Reference Sans
Serif');
% Create ylabel
ylabel('PACF koeficienty (-)  \rightarrow',...
    'FontName','MS Reference Sans Serif');
% Create title
title('PACF {\it} ',...
    'FontName','MS Reference Sans Serif');


figure(2)

subplot(2,1,1), bar(dacf);
% Create xlabel
xlabel('{\itposunutie } (-)  \rightarrow','FontName','MS Reference Sans
Serif');
% Create ylabel
ylabel('ACF koeficienty(-)  \rightarrow',...
    'FontName','MS Reference Sans Serif');
% Create title
title('ACF po 1. diferenciácii ČR {\it} ',...
    'FontName','MS Reference Sans Serif');

subplot(2,1,2), bar(dpacf);

% Create xlabel
xlabel('{\itlag } (-)  \rightarrow','FontName','MS Reference Sans
Serif');
% Create ylabel
ylabel('PACF koeficienty (-)  \rightarrow',...
    'FontName','MS Reference Sans Serif');
% Create title
title('PACF po 1. diferenciácii ČR {\it} ',...
    'FontName','MS Reference Sans Serif');

figure(3)

subplot(2,1,1), stem(trainSeries,'k','Marker','none','LineWidth',2);

% Create xlabel
xlabel('{\itt } (30 s)  \rightarrow','FontName','MS Reference Sans
Serif');
% Create ylabel
ylabel('HTTP požiadavky (-)  \rightarrow',...
    'FontName','MS Reference Sans Serif');
% Create title
title('Pôvodný časový po 1. diferenciácii {\it} ',...
    'FontName','MS Reference Sans Serif');

subplot(2,1,2), stem(simSeries ,'r','Marker','none','LineWidth',2);
% Create xlabel
```

```
xlabel('{\itt } (30 s)  \rightarrow','FontName','MS Reference Sans
Serif');
% Create ylabel
ylabel('HTTP požiadavky (-)  \rightarrow',...
    'FontName','MS Reference Sans Serif');
% Create title
title('Simulovaný časový rad podľa modelu {\it} ',...
    'FontName','MS Reference Sans Serif');

figure(4)

subplot(2,1,1), stem(fullSeries(5701:5760,1),'k','Marker','none',...
    'LineWidth',2);

% Create xlabel
xlabel('{\itt } (30 s)  \rightarrow','FontName','MS Reference Sans
Serif');
% Create ylabel
ylabel('HTTP požiadavky (-)  \rightarrow',...
    'FontName','MS Reference Sans Serif');
% Create title
title('Vynechané hodnoty pôvodného ČR{\it} ',...
    'FontName','MS Reference Sans Serif');

subplot(2,1,2), stem((SigmaForecast +
MeanForecast),'r','Marker','none',...
    'LineWidth',2);
% Create xlabel
xlabel('{\itt } (30 s)  \rightarrow','FontName','MS Reference Sans
Serif');
% Create ylabel
ylabel('HTTP požiadavky (-)  \rightarrow',...
    'FontName','MS Reference Sans Serif');
% Create title
title('Predikované hodnoty {\it} ',...
    'FontName','MS Reference Sans Serif');
```